Fast Speech in Unit Selection Speech Synthesis

Dissertation
zur Erlangung des Grades eines Dr. phil.

der Fakultät für Linguistik und Literaturwissenschaft
der Universität Bielefeld vorgelegt von

Donata Jadwiga Moers-Prinz

aus Jülich

Gutachter:
Prof. Dr. Petra S. Wagner
Prof. Dr. Bernd Möbius

Tag der mündlichen Prüfung:
08. Februar 2019

Eigenständigkeitserklärung

Ich, Donata Jadwiga Moers-Prinz, versichere, dass ich die Dissertation ”Fast Speech in Unit Selection Speech Synthesis” selbständig verfasst habe. Ich versichere, dass

- mir die Promotionsordnung der Fakultät für Linguistik und Literaturwissenschaft der Universität Bielefeld bekannt ist;
- ich die Dissertation selbst angefertigt habe, keine Textabschnitte von Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen in dieser Arbeit angegeben habe;
- Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Vermittlungstätigkeiten oder Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;
- ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;
- ich weder diese Dissertation, noch eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Ort, Datum   Unterschrift

Gedruckt auf alterungsbeständigem Papier °°ISO 9706

**Whom have we conquered?
None but ourselves.**

**Have we won a kingdom?
No - and yes.
We have achieved an ultimate satisfaction,
fulfilled a destiny.**

George Mallory, 1918 (after Charles S. Houston, 1968)

Zusammenfassung

Forschungsvorhaben im Bereich der Sprachsynthese können unterschiedlich motiviert sein. Oft stehen der Sprachproduktionsprozess und die akustischen Eigenschaften von Sprache im Mittelpunkt des Interesses, aber auch die Entwicklung einer möglichst optimalen, natürlich klingenden Sprachsynthese ist ein häufiges Forschungsziel. In Abhängigkeit von diesem Ziel spielt auch die Wahl des Syntheseverfahrens eine wichtige Rolle: Artikulatorische Synthese wird eher zur Untersuchung des Sprachproduktionsprozesses und insbesondere artikulatorischer Prozesse eingesetzt. Konkatenative, auf Einheiten-Auswahl basierende Sprachsynthese wird hingegen bevorzugt dann verwendet, wenn es um die Erzeugung möglichst natürlich klingender Sprache geht, da bei dieser Sprachsynthesetechnologie Einheiten, die aus natürlicher Sprache ausgeschnitten werden, zu neuen Phrasen und Äußerungen zusammengesetzt werden.

Insbesondere blinde und sehbehinderte Mitmenschen nutzen Sprachsynthese heutzutage häufig als IT-Hilfsmittel. Oft bevorzugen sie hierbei im Rahmen der täglichen Verwendung dieses Hilfsmittels eine hohe bis sehr hohe Sprechrate [Granström, 1991], [Fellbaum, 1996], [Portele and Krämer, 1996], [He and Gupta, 2001], [Asakawa et al., 2002], [Nishimoto et al., 2006], [Moos and Trouvain, 2007], [Borodin et al., 2010], [Ahmed et al., 2012], [McCarthy et al., 2013]. In Einheiten-Auswahl-basierten Sprachsynthesesystemen, die mit Bausteinen aus natürlicher Sprache arbeiten, wird eine hohe Sprechgeschwindigkeit aber nur unzureichend modelliert. Sprachsynthesesysteme, die mit Signalmanipulation arbeiten, können zwar hohe Sprechraten erzeugen, die Manipulation entspricht aber nicht den tatsächlich in natürlicher, schnell gesprochener Sprache zu beobachtenden Phänomenen. Darüberhinaus geht bei Verwendung eines solchen Systems häufig die Natürlichkeit der erzeugten Sprache verloren. Ziel der hier vorgestellten Arbeit ist es, die Möglichkeiten zur Modellierung hoher Sprechgeschwindigkeit in der auf Einheiten-Auswahl basierenden Sprachsynthese zu untersuchen. Auf diesem Wege wird eine Strategie zur Erzeugung und Integration dieses Sprechstils in diese Art von Sprachsynthesesystemen entwickelt.

Um schnelle Sprache in der Einheiten-Auswahl-basierten Sprachsynthese modellieren zu können, mussten zunächst einmal die typischen phonetischen Eigenschaften schnell gesprochener Sprache analysiert werden. Die phonetischen Merkmale schnell gesprochener Sprache unterscheiden sich stark von den Eigenschaften normal gesprochener Sprache, daher werden sie in Kapitel 2 umfassend erläutert. Zunächst sind dabei der Begriff der Sprechgeschwindigkeit und ihre Quantifizierung von besonderem Interesse (vgl. Abschnitt 2.1). Die allgemein zu beobachtenden Phänomene Koartikulation, Reduktion und Elision werden anschließend in Abschnitt 2.2.1 betrachtet. Die Veränderung der Sprechgeschwindigkeit beeinflusst aber nicht nur den Artikulationsvorgang als

solchen (vgl. Kapitel 2.2.2), sondern auch die akustischen Eigenschaften einzelner Laute und Lautübergänge, wie in den Kapiteln 2.2.2 und 2.2.2 beschrieben. Darüber hinaus sind auch suprasegmentale Merkmale wie Betonung, Grundfrequenzverlauf und Phrasierung, deren Betrachtung sich in Abschnitt 2.2.3 wiederfindet, von Änderungen der Sprechgeschwindigkeit betroffen. Der letzte Abschnitt des ersten Kapitels, Abschnitt 2.3, befasst sich mit verschiedenen Strategien der Sprachproduktion sowie der Realisierung unterschiedlicher Sprechstile. Hier wird der für die Durchführung der vorgesehenen Integration schneller Sprache in die Einheiten-Auswahl-basierte Sprachsynthese erforderliche Sprechstil definiert: Die schnell gesprochene Sprache soll so deutlich und präzise wie möglich artikuliert werden, um unerwünschte Phänomene wie Koartikulation und Reduktion, die die Verständlichkeit schnell gesprochener Sprache im Allgemeinen negativ beeinflussen, weitestgehend zu vermeiden.

Da die Unterscheidung zwischen verschiedenen Methoden der Sprachsynthese mit Blick auf das Forschungsziel eine wichtige Rolle spielt, werden anschließend in Kapitel 3.1 verschiedene Arten der Sprachsynthese wie auch deren Vor- und Nachteile dargestellt. Es wird erläutert, dass die Einheiten-Auswahl-basierte Sprachsynthese für die Durchführung des hier vorgestellten Forschungsvorhabens als geeignete Technologie angesehen wurde, da sie am besten für die Erzeugung natürlich klingender Sprache geeignet ist (siehe Abschnitt 3.1.2). Nichtsdestotrotz wird auch verdeutlicht, dass diese Art der künstlichen Spracherzeugung einen entscheidenden Nachteil mit sich bringt: Die glatten Lautübergänge, die aufgrund koartikulatorischer Phänomene bei der Produktion natürlicher Sprache zu beobachten sind, sind für die Verständlichkeit natürlicher ebenso wie künstlich erzeugter Sprache von besonderer Bedeutung [Martinez et al., 1997], [Winters and Pisoni, 2004]. Diese werden aber bei der auf Einheiten-Auswahl basierenden Sprachsynthese häufig durch die Konkatenation von Einheiten verschiedener Herkunft zerstört. Von parametrischer Sprachsynthese hingegen (siehe Abschnitt 3.1.1), bei der das Sprachsignal komplett künstlich erzeugt wird, werden derartige Lautübergänge sehr gut reproduziert. Darüberhinaus erlauben die in parametrischer Synthese verwendeten Modelle eine flexible Anpassung an andere Sprechstile, sofern sie entsprechend trainiert wurden, wohingegen der Sprechstil der mit Einheiten-Auswahl-basierten Sprachsynthese erzeugten künstlichen Sprache vom Sprechstil der zugrundeliegenden Korpusaufnahmen abhängig ist [Zen et al., 2007]. Die zeitaufwändige Erstellung eines zusätzlichen, schnell gesprochenen Bausteininventars macht somit nur dann Sinn, wenn deutliche perzeptive Unterschiede zum bisher verwendeten Bausteininventar zu verzeichnen sind.

In der vorliegende Arbeit wurden zwei bestimmte Sprachsynthesesysteme gegenübergestellt, die die beiden oben dargelegten, unterschiedlichen Herangehensweisen an die Problemstellung widerspiegeln: Zum einen ist dies die als IT-Hilfsmittel weit verbreitete, parametrische Sprachsynthese "JAWS Eloquence" [FreedomScientific, 2011], zum anderen das Einheiten-Auswahl-basierte Sprachsynthesesystem "BOSS" (Bonn Open Speech Synthesis, [Klabbers et al., 2001]).

Die Systemarchitektur von BOSS wurde abschließend in Kapitel 3.1.2 näher erläutert, da für die Implementierung schnell gesprochener Sprache gewisse Anpassungen einzelner Komponenten des Systems erforderlich waren (siehe Abschnitt 7.2).

In Kapitel 3.2 wurden verschiedene Ansätze und Modelle zur Modellierung der Sprechgeschwindigkeit in der Sprachsynthese aufgezeigt. Zu Anfang wurden in Abschnitt 3.2.1 unterschiedliche Vorgehenweisen zur Vorhersage der zu generierenden Einheitendauer näher erläutert. Eine adäquate Dauervorhersage ist für die wahrgenommene Natürlichkeit künstlich generierter Sprache von besonderer Bedeutung [Brinckmann and Trouvain, 2003]. Da die Dauer sprachlicher Einheiten von sehr vielen verschiedenen Faktoren beeinflusst wird, wurde angenommen, dass die Implementierung schnell gesprochener Sprache in das Einheiten-Auswahl-basierte Sprachsynthesesystem BOSS eine Anpassung der Modelle zur Dauervorhersage erforderlich machen würde. Inwiefern diese Annahme zutreffend war, wurde anhand des weitverbreiteten Ansatzes der Erstellung so genannter "Classification And Regression Trees" (CART, [Breiman et al., 1984]) überprüft. Da diese Vorgehensweise keine Verwendung von manuell erzeugten Regeln zur Dauervorhersage erfordert und auch in der Lage ist, größere Datenmengen problemlos zu verarbeiten, wurde sie hier als vielversprechende Methode zur Vorhersage der Dauer einzelner Spracheinheiten in den verschiedenen Sprechstilen angesehen.

Um schnelle Sprache in der Einheiten-Auswahl-basierten Sprachsynthese zu modellieren, gab es bisher nur die Möglichkeit, die in normaler Sprechgeschwindigkeit erzeugten Sprache mit Hilfe von Dauermanipulation linear oder nicht-linear zu beschleunigen, wobei ggf. verschiedene prosodische Eigenschaften schnell gesprochener Sprache, wie z. Bsp. die Pausendauer, der Intonationsverlauf und die Stärke der Phrasengrenzen, imitiert wurden [Trouvain, 2002a], [Trouvain, 2002b]. Am weitesten verbreitet ist heutzutage allerdings die lineare Dauermanipulation, die mit Hilfe des so genannten Pitch-synchronous overlap add (PSOLA; [Moulines and Charpentier, 1990], [Liu and Zeng, 2006]) -Verfahrens durchgeführt wird. Dieses hat jedoch den Nachteil, dass bei einer Beschleunigung des Signals um den Faktor zwei oder mehr störende Artefakte entstehen. Eine andere, neuere Möglichkeit zur Generierung schnell gesprochener Sprache in der auf Einheiten-Auswahl basierenden Sprachsynthese, die in der vorliegenden Arbeit evaluiert wird, ist die Erstellung eines natürlichsprachlichen Bausteininventars, welches alle segmentalen und suprasegmentalen Eigenschaften schnell gesprochener Sprache bereits beinhaltet. Vorhergehende Studien haben gezeigt (z. Bsp. [Janse, 2003b]), dass diese Vorgehensweise zu einer veränderten Wahrnehmung der schnell gesprochenen Sprache führt. Insbesonders ausnehmend deutlich und präzise schnell gesprochene Sprache lässt einen perzeptiven Vorteil gegenüber unpräzise artikulierter, schneller Sprache erwarten. Nichtsdestotrotz ist die Anwendung von PSOLA für die Erzeugung ultra-schneller Sprache, deren Sprechgeschwindigkeit über die natürlich produzierbare, schnelle Sprechgeschwindigkeit deutlich hinausgeht (siehe Kapitel

2.1, [Moos and Trouvain, 2007]), wie es von einigen trainierten Benutzern von Sprachsynthesesystemen bevorzugt wird [Portele and Krämer, 1996], [Moos and Trouvain, 2007], unverzichtbar.

Von den zu beobachtenden akustischen Veränderungen schnell gesprochener Sprache wird auch die Perzeption beeinflusst. Daher werden in Kapitel 4 verschiedene Aspekte der Wahrnehmung schneller natürlicher und künstlich erzeugter Sprache diskutiert. In Abschnitt 4.1 werden zunächst allgemeine Aspekte der Perzeption natürlicher Sprache dargelegt. Verschiedene Modelle, mit denen Sprachwahrnemung auf abstrakter Ebene beschrieben wird, werden anschließend in Kapitel 4.1.1 dargelegt. Danach liegt der Fokus der Ausführungen auf Mechanismen, die bei der perzeptiven Anpassung an veränderte spektrale oder zeitliche Eigenschaften von natürlicher Sprache zu beobachten sind. Analog zu den Ausführungen zur Messung der Sprechgeschwindigkeit in Kapitel 2.1 schließt der erste Teil dieses Kapitels mit Erläuterungen zu den Einheiten der Wahrnehmung von Sprechgeschwindigkeit (siehe Abschnitt 4.1.2).

Im Anschluss wird in Abschnitt 4.2 die Wahrnehmung künstlich erzeugter schneller Sprache näher betrachtet. Allgemein anerkannte Methoden zur Evaluierung künstlich erzeugter Sprache stehen dabei zunächst in Abschnitt 4.2.1 im Mittelpunkt. Die Methoden, die in der vorliegenden Arbeit zur Evaluierung der erzeugten schnell gesprochenen Sprache zur Anwendung kommen, werden hier definiert. Neben der Beurteilung der Verständlichkeit und Natürlichkeit, welche auf so genannten Mean Opinion Scores (MOS) beruht, wurde die Word Error Rate (WER) gewählt, um die Perzeption der unter verschiedenen Voraussetzungen generierten schnellen Sprache zu evaluieren. Danach konzentrieren sich die Darlegungen auf die Wahrnehmung von beschleunigter sowie künstlich erzeugter Sprache (vgl. Abschnitte 4.2.2 und 4.2.3). Verschiedene Untersuchungen haben gezeigt, dass die Perzeption beschleunigter und synthetisierter (schneller) Sprache für den Hörer schwieriger ist als die Wahrnehmung von in normalem Tempo gesprochener, natürlicher Sprache [Winters and Pisoni, 2004], [Papadopoulos et al., 2010]. [Winters and Pisoni, 2004] schließen ihre Ausführungen allerdings mit der Anmerkung, dass die kognitive Verarbeitung synthetischer Sprache länger dauern mag als die Verarbeitung natürlicher Sprache, dass aber ihrer Meinung nach die erreichbare endgültige Stufe des Verstehens bei beiden gleich sei. [Schwab et al., 1985] beobachteten darüber hinaus einmal mehr, dass insbesondere die in konkatenierter, synthetischer Sprache auftretenden Diskontinuitäten zur Beeinträchtigung der Wahrnehmung beitrugen. Andererseits war gerade die Beinhaltung robuster, redundanter perzeptiver Merkmale einzelner Segmente ein Vorteil der konkatenativen Einheiten-Auswahl-Synthese. Diese Überlegungen zur Wahrnehmung beschleunigter und synthetisierter Sprache sind auch in die Erhebungen der vorliegenden Arbeit eingeflossen: Zum einen bei der Evaluierung der natürlichen, schnell gesprochenen Sprache der ausgewählten Sprecherin im Vergleich zu ihrer beschleunigten, in normalem Tempo gesprochenen Sprache, wie in

Kapitel 7.1.1 beschrieben, und zum anderen im Zusammenhang mit der Evaluierung der mit verschiedenen Sprachsynthesesystemen und unterschiedlichen Korpora generierten (ultra-)schnellen Sprache (vgl. Abschnitt 8.2).

Zum Schluss des Kapitels wird die Beeinflussung der Wahrnehmung synthetisierter Sprache und somit auch deren perzeptiver Beurteilung durch Hörergegebenheiten betrachtet (siehe Abschnitt 4.3). Neben individuellen, physiologischen Eigenschaften der Hörer spielt hier insbesondere die langfristige und regelmäßige Nutzung bestimmter Sprachsynthesetechnologien eine wichtige Rolle. Die hier dargestellten Untersuchungen weisen eindeutig darauf hin, dass über die Zeit eine perzeptive Anpassung an eine bestimmte Art von schnell gesprochener Sprache stattfindet, und somit ein gewisser Trainingseffekt zu verzeichnen ist, auch wenn die Adaption im Falle von künstlich erzeugter Spache einen längeren Zeitraum in Anspruch nimmt als bei natürlicher Sprache. Inwiefern diese Anpassung auch die Wahrnehmung von mit Einheiten-Auswahl-basierter Sprachsynthese erzeugter Sprache beeinflusst, wird später in Abschnitt 8.2 ausführlich diskutiert.

Während der Vorbereitung der empirischen Studien kamen einige grundlegende Fragen bzgl. der Wünsche und Gewohnheiten der möglichen Hauptnutzer des neu zu entwickelnden, Einheiten-Auswahl-basierten Sprachsynthese-Sprechstils auf: Welche Qualität von synthetischer Sprache wünschten sich die Nutzer von Sprachsynthese als IT-Hilfsmittel generell (vgl. [Granström, 1991], [Fellbaum, 1996], [Brinckmann and Trouvain, 2003], [Stent et al., 2011])? War es richtig, dass eine monotone, prosodisch flache Sprachausgabe bevorzugt wurd, wie von [Fellbaum, 1996] behauptet? War es den blinden und sehbehinderten Nutzern von Sprachsynthese egal, wenn die Sprachausgabe nicht natürlich klang, so lange sie verständlich war und insbesondere Lautübergänge, die für die Verständlichkeit so wichtig sind, gut modellierte, wie in parametrischer Synthese (vgl. [Moos and Trouvain, 2007])? Die Untersuchungen von [Portele and Krämer, 1996] und [McCarthy et al., 2013] wiesen jedenfalls darauf hin, dass für die meisten Nutzer die Einfachheit der Anwendung, die Flexibilität des Systems sowie dessen Robustheit mindestens genauso wichtige Kriterien waren wie die Natürlichkeit der Sprachausgabe. Während neue Nutzer noch Wert auf eine natürlich klingende Stimme und generelle Stimmqualität legten, waren für fortgeschrittene Benutzer der technische Support sowie die Möglichkeit zur übergangslosen Beschleunigung der Sprechgeschwindigkeit von höchster Wichtigkeit [Chalamandaris et al., 2010], [McCarthy et al., 2013]. Die Studie von [Chalamandaris et al., 2010] brachte sogar die Forderung nach der Möglichkeit, die Stimm- bzw. Sprachausgabequalität für schnelle Sprache zu reduzieren hervor. Um zu vermeiden, an den Bedürfnissen der Hauptnutzer vorbei zu forschen [Wagner, 2013], wurde daher zunächst eine Online-Umfrage durchgeführt, um diese Fragen zu klären [Moers et al., 2007]. Details und Ergebnisse der Umfrage wurden in Kapitel 5 dargelegt. Es zeigte sich, dass geübte blinde und sehbehinderte Nutzer von Sprachsynthese als IT-Hilfsmittel tatsächlich eine schnelle

Sprechgeschwindigkeit bevorzugten, wenn auch nicht alle in gleichem Maße. Allerdings wiesen die Umfrageergebnisse auch darauf hin, dass deutlich weniger als die Hälfte der Nutzer eine monotone Intonation bevorzugten oder gar dazu bereit waren, auf jegliche Art der Intonation zu verzichten. Nichtsdestotrotz wurde von allen Umfrageteilnehmern die Verständlichkeit als wichtigstes Qualitätskriterium benannt, wohingegen Natürlichkeit von einem Drittel der Befragten als unwichtig betrachtet wurde [Moers et al., 2007]. Viele Teilnehmer wiesen darauf hin, dass insbesondere die bei konkatenativer Sprachsynthese auftretenden Störungen des Sprachsignals bislang so gravierend seien, dass die deutlich höhere Natürlichkeit als Beurteilungskriterium keine Rolle mehr spiele. Die Aussagen zur gewünschten Natürlichkeit und Lebhaftigkeit wurden als Bestätigung des Vorhabens angesehen, im Rahmen der vorliegenden Arbeit robuste Richtlinien zur Implementierung schnell gesprochener Sprache in die Einheiten-Auswahl-basierte Sprachsynthese zu entwickeln. Darüber hinaus lieferten die Anmerkungen zur negativen Auswirkung der hohen Anzahl von Konkatenationsstellen den Ausgangspunkt für die Überlegung, eine passendere Einheiten-Definition als die bisher übliche zu finden. Die zu diesem Aspekt durchgeführten Analysen werden in Kapitel 8.1 näher erläutert.

Der Sprecher des Baustein-Inventars für die Einheiten-Auswahl-basierte Sprachsynthese bestimmt zu einem großen Teil die spätere Qualität der synthetisierten Sprache [Syrdal et al., 1997]. Von dem oben beschriebenen erforderlichen Sprechstil (schnell und deutlich) wurden daher grundlegende Anforderungen an die Sprecherin bzw. den Sprecher abgeleitet. So sollte sie/er vor allem in der Lage sein, bei möglichst schnellem Sprechtempo immer noch sehr deutlich zu artikulieren. Die genaue Vorgehensweise bei der Auswahl der Sprecherin bzw. des Sprechers wurde zu Beginn von Kapitel 6 beschrieben. In Abschnitt 6.2.1 wurden dann die akustischen Eigenschaften der deutlich artikulierten schnellen Sprache der ausgewählten Sprecherin mit ihrer weniger deutlich artikulierten schnellen Sprache verglichen und anschließend in Relation gesetzt zu ihrer in normalem Tempo produzierten Sprache. Abschnitt 6.2.2 beschreibt im Anschluss die perzeptive Evaluation der verschiedenen Sprechstile. Hatte die akustische Analyse verschiedener Merkmale der von der ausgewählten Sprecherin produzierten Sprechstil-Varianten noch keine eindeutigen Hinweise auf die Erfüllung der aufgestellten Eignungskriterien geliefert, zeigte die perzeptive Evaluation eine eindeutige Präferenz der Hörer bgzl. der schnell und deutlich artikulierten Sprechvariante. Hieraus wurde geschlossen, dass die gewählte Sprecherin in der Tat für die Erstellung eines schnell und deutlich gesprochenen Bausteininventars für die Einheiten-Auswahl-basierte Sprachsynthese geeignet war.

Um die Modellierung schnell gesprochener Sprache in der auf Einheiten-Auswahl basierenden Sprachsynthese zu untersuchen, werden danach zwei voneinander unabhängige, jedoch inhaltlich identische Inventare zur Einheiten-Auswahl erstellt: Eines in normaler Sprechgeschwindigkeit und eines in möglichst schneller und deutlicher Sprache. Die Arbeitsschritte, die hierbei durchgeführt wurden, wurden in Kapitel 7 ausführlich beschrieben. Um ein Synthese-Inventar zu entwickeln, das sowohl leicht handhabbar als auch gut nutzbar sein würde, wurde dann erneut untersucht, ob die von der Sprecherin erzeugte schnell und deutlich gesprochene Sprache einen perzeptiven Nachteil gegenüber der in normalem Tempo produzierten Sprache aufweisen würde. Die Details dieser Analyse wurden in Kapitel 7.1.1 dargelegt. Anschließend wurden Implikationen für die Implementierung schnell gesprochener Sprache als separates Inventar diskutiert.

Da es ein erklärtes Ziel der vorliegenden Arbeit war, robuste Richtlinien für die Integration schnell gesprochener Sprache in die Einheiten-Auswahl-basierte Sprachsynthese zu erstellen, wurden im weiteren Verlauf Verfahren der Korpusaufbereitung auf die schnell gesprochene Sprache angewendet, die üblicherweise auch bei der Implementierung eines Korpus in normal gesprochener Sprache zum Einsatz kommen. Das o.a. CART-Verfahren zur Vorhersage der Segmentdauer war eines dieser Verfahren. Es wurden separate Modelle für die Vorhersage der Segmentdauern in normaler sowie in schneller, deutlicher Sprache mittels CART erstellt. Wichtige phonetische und prosodische Faktoren, die im Allgemeinen einen Einfluss auf die Segmentdauer haben, wurden hierbei berücksichtigt. Die Ergebnisse der Anwendung von CART auf die normale wie auch auf die schnell und deutlich gesprochene Sprache zeigten, dass die Korrelation zwischen der beobachteten und der vorhergesagten Dauer für beide Sprechgeschwindigkeiten vergleichbar war (siehe Abschnitt 7.2.2). Daraus wurde geschlossen, dass CART-basierte Dauerprädiktion auch für die Vorhersage der Dauer von Sprachsegmenten in schnell und deutlich gesprochener Sprache anwendbar sein würde.

Die Aufbereitung der Korpusaufnahmen zum Einheiten-Auswahl-Inventar ist eine der aufwändigsten Aufgaben bei der Implementierung eines neuen Sprechstils in diese Art der Sprachsynthese. Insbesondere das häufig erforderliche manuelle Korrigieren von Segmentierungs- bzw. Label-Grenzen benötigt extrem viel Zeit. Aus diesem Grunde kommen hier bevorzugt automatische Verfahren zum Einsatz. Da die Qualität der erzeugten Sprache jedoch auch von der so genannten Label Timing Accuracy (LTA) abhängt [Kominek et al., 2003], war es fragwürdig, ob die Anwendung eines solchen Verfahrens für schnell und deutlich gesprochene Sprache ebenfalls geeignet sein würde, da für diesen Sprechstil eine erhöhte Anzahl falsch gesetzter Segmentgrenzen zu erwarten war. Sollte sich diese Annahme als wahr erweisen, wäre dieses automatische Verfahren nicht auf schnell gesprochene Sprache anwendbar und somit eine Implementierung dieses Sprechstils in die Einheiten-Auswahl-basierte Sprachsynthese aus rein praktischen Gründen nicht empfehlenswert (siehe auch [Wagner, 2013]). In Abschnitt 7.2.1 findet sich eine detaillierte Analyse der Label Timing Accuracy sowohl für die normale als auch für die schnell und deutliche artikulierte Sprache der gewählten Sprecherin. Die zur Vorbereitung der Verarbeitung erforderlichen Arbeitsschritte umfassten dabei eine Anpassung der vorhandenen Transkriptionen an das Sprachsynthesesystem BOSS und die automatische Segmentierung beider Korpora mit Hilfe eines HTK-basierten Aligners, der an Deutsch angepasst wurde [Dragon, 2005]. Eine Analyse der LTA nach Durchführung der automatischen Segmentierung beider Korpora zeigte nur marginale Unterschiede zwischen den beiden Sprechgeschwindigkeitsvarianten. Daraus wurde gefolgert, dass automatische Segmentierungsverfahren auch auf schnell und deutlich gesprochene Sprache anwendbar sind. Nichtsdestotrotz wären mit Blick auf die schnell und deutlich gesprochene Sprache möglicherweise eindeutigere Ergebnisse erzielt worden, wenn das gemessene Toleranzintervall an die Gegebenheiten schneller Sprache - im Sinne von kürzerer durchschnittlicher Dauer der in ihr enthaltenen Sprachsegmente - angepasst worden wäre.

Wie zu Beginn bereits festgestellt, bevorzugen blinde und sehbehinderte Nutzer von Sprachsynthese als IT-Hilfsmittel häufig die weniger natürlich klingende parametrische Synthese, da diese es ermöglicht, Sprache, und insbesondere einzelne Lautübergänge, ohne störende Konkatenationsphänomene zu generieren. Da Diskontinuitäten für konkatenative Synthese im Allgemeinen und Einheiten-Auswahl-basierte Sprachsynthese im Besonderen ein Problem darstellen, haben [Breuer and Abresch, 2004] vorgeschlagen, bestimmte Phonem-Sequenzen, bei denen starke Koartikulationseffekte auftreten, zwar als zwei oder mehr Phone zu betrachten, bei ihrer Verwendung zur Erzeugung synthetischer Sprache aber als unzertrennliche Syntheseeinheiten zu behandeln - so genannte Phoxsy-Einheiten [Breuer and Abresch, 2004]. Dieser Ansatz könnte zu einer möglichen Lösung für die Modellierung schnell gesprochener Sprache in der Einheiten-Auswahl-basierten Sprachsynthese führen, da durch die Verwendung aufgezeichneter, natürlich-sprachlicher Bausteine sowohl die Natürlichkeit erhalten, als auch die Verständlichkeit durch die Berücksichtigung der Lautübergänge bei stark koartikulierten Lautkombinationen möglichst hoch bleibt [Winters and Pisoni, 2004]. Der nächste Schritt war somit die Durchführung einer Analyse, ob dieser Ansatz zur Definition der Einheitengröße auch bei schnell und deutlich gesprochener Sprache Anwendung finden könnte. Es wurde erwartet, durch die Verwendung von Phoxsy-Einheiten eine Möglichkeit zu finden, schnelle Sprache in der Einheiten-Auswahl-basierten Sprachsynthese besser modellieren zu können. Dabei würde die Art der Sprachsynthese die Natürlichkeit der generierten schnellen Sprache erhöhen, wohingegen die Verwendung größerer Syntheseeinheiten der allgemeinen Verständlichkeit zuträglich wäre. Die Methoden und Ergebnisse dieser Evaluierung sind in Kapitel 8.1 dargelegt: Phoxsy-Einheiten wurden für beide Sprechgeschwindigkeiten in das Sprachsynthesesystem BOSS [Klabbers et al., 2001] integriert. Für Sprache in normalem Sprechtempo konnten die Ergebnisse von [Breuer and Abresch, 2004] bestätigt werden: Stimuli, die unter alleiniger Verwendung von Phoxsy units generiert wurden, wurden als signifikant besser verständlich beurteilt als Stimuli, die nur aus einzelnen Phonen generiert wurden. Dies galt gleichermaßen für Stimuli, die unter Verwendung aller Einheiten-Auswahl-Ebenen inklusive Phoxsy units generiert wurden. Die Natürlichkeit wurde von der Wahl der zur Synthese genutzten Einheitengröße jedoch nicht signifikant beeinflusst. Ein deutlich anderes Bild zeigte sich in Bezug auf die synthetisierte schnell gesprochene Sprache: Hier führte die Verwendung von Phoxsy units zu einer signifikant besseren Beurteilung von Verständlichkeit und Natürlichkeit der Stimuli. Aus diesen Ergebnissen wurde geschlossen, dass Phoxsy units nicht nur zur Verbesserung der Verständlichkeit schnell gesprochener Sprache geeignet waren, sondern auch wesentlich zu einer Erhöhung der wahrgenommenen Natürlichkeit beitragen konnten.

Nachdem die vorherigen Experimente gezeigt hatten, dass Phoxsy units zur Generierung schnell gesprochener Sprache besser geignet waren als herkömmliche Einheiten, wurden die Verständlichkeit, die Natürlichkeit und die allgemeine Annehmbarkeit von schnell gesprochenen Äußerungen genauer untersucht. Diese wurden mit Hilfe unterschiedlicher Syntheseverfahren und - bei der Einheiten-Auswahl-basierten Sprachsynthese - unterschiedlicher Korpora in verschiedenen Sprechgeschwindigkeiten generiert. Hierzu wurden zunächst mehrere Gruppen so genannter Semantisch Unvorhersagbarer Sätze (Semantically Unpredictable Sentences, SUS, [Benoit and Grice, 1996]) erzeugt [Benoit and Grice, 1996] (siehe auch [Syrdal et al., 2012]). Anschließend wurde von zwei unterschiedlichen Hörergruppen ein Mean Opinion Score (MOS) für die so generierten Äußerungen erhoben. Darüber hinaus wurde die Anzahl der verstandenen Inhaltswörter erfasst, um darauf basierend später die Wortfehlerrate (Word Error Rate, WER) in Abhängigkeit von der Hörergruppe, dem Synthesesystem und der Sprechgeschwindigkeit zu ermitteln. Daneben wurde als weitere Variable die Anzahl der WIederholungen beim Anhören der jeweilgen Äußerung notiert. Die eine Hörergruppe bestand aus geübten, meist blinden oder sehbehinderten Hörern, die seit mehr als zwei Jahren nahezu täglich einen Screenreader, basierend auf dem Formantsynthesesystem "JAWS Eloquence"[FreedomScientific, 2011] benutzten, die anderen Gruppe bestand aus sehenden oder erst kürzlich erblindeten Probanden, die bis dato kaum oder gar keine Erfahrung mit einer solchen Anwendung gesammelt hatten. Die Ergebnisse der statistischen Analyse wurden in Kapitel 8.2 mit Hilfe von binären Entscheidungsbäumen dargestellt. Es zeigte sich, dass die Verständlichkeit das wichtigste Beurteilungskriterium war, in übereinstimmung mit [Stent et al., 2011] und [McCarthy et al., 2013]. Weder Natürlichkeit noch Stimmqualität spielten bei der Vergabe des MOS eine große Rolle. Darüber hinaus wurde deutlich, dass die geübten Hörer der ersten Hörergruppe die Formantsynthese gegenüber der Einheiten-Auswahl-basierten Sprachsynthese deutlich bevorzugten (vgl. [Winters and Pisoni, 2004]). Daraus ließ sich ableiten, dass diese Hörergruppe durch die dauerhafte Nutzung des o.a. Formantsynthesesystems als IT-Hilfsmittel voreingenommen und darauf trainiert war. Bei der Gruppe der unerfahrenen Hörer war das den Stimuli zugrundeliegende Synthesesystem hingegen unwichtig für die Vergabe des MOS; vielmehr war hier die Sprechgeschwindigkeit des zu beurteilenden Stimulus das wichtigste Beurteilungskriterium. Erst nachdem diese zwei Hauptkriterien bei der Auswertung beiseite gelassen wurden, um weitere Unterschiede zu entdecken, wurde anhand des neu berechneten Entscheidungsbaumes deutlich, dass sowohl die Sprechgeschwindigkeit als auch das jeweilige Einheiten-Auswahl-Inventar sowie die Zugehörigkeit zu einer der beiden Hörergruppen eine wichtige Rolle bei der Beurteilung der verschiedenen Stimuli spielten. Stimuli, die auf dem in normaler Sprechgeschwindigkeit gesprochenen Bausteininventar basierten, wurden von beiden Hörergruppen als signifikant besser bewertet als die Stimuli, die auf dem schnell gesprochenen Bausteininventar basierten. Dessen ungeachtet unterschieden ungeübte Hörer bei extrem hoher Sprechgeschwindigkeit jedoch nicht zwischen Einheiten-Auswahl-basierter Sprachsynthese und Formantsynthese, im Gegensatz zu geübten Hörern, bei denen die gewohnte Formantsynthese besser abschnitt als die Einheiten-Auswahl-basierte Sprachsynthese.

Die zwei wichtigsten Erkenntnisse der abschließenden Analyse sind somit zum einen die Bestätigung des Vorhandenseins eines Trainingseffekts für eine bestimmte Art der Sprachsynthese, welcher aber erst nach längerer Exposition auftritt und sich dann insbesondere in der signifikant besseren Worterkennungsrate sowie der signifikant besseren allgemeinen Beurteilung des betroffenen Sprachsynthesesystems widerspiegelt. Zum anderen wurde trotz der vielversprechenden Zwischenergebnisse bzgl. der erfragten akustisch-perzeptiven Präferenzen der Nutzer sowie der Verarbeitbarkeit und Verwendbarkeit eines schnell gesprochenen Einheiten-Auswahl-Inventars deutlich, dass der gewählte Ansatz zur Modellierung dieses Sprechstils in der Einheiten-Auswahl-basierten Sprachsynthese keine Vorteile hinsichtlich der Natürlichkeit, der Verständlichkeit und der allgemeinen Annehmbarkeit mit sich bringt, insbesondere dann nicht, wenn die angestrebte Sprechgeschwindigkeit der Sprachausgabe die in natürlicher, schnell gesprochener Sprache erreichbare Sprechgeschwindigkeit deutlich übersteigt.

Acknowledgements

To finalize this work has taken quite some time, and quite some people crossed my way during this time. To list all of you might need too much time and space, so I would simply like to say ”thank you” to all of you. However, I would also like to send some more detailed words of thanks to those who in one way or another contributed to this work which finally has come to an end now:

First of all, I would like to thank Prof. i.R. Dr. Wolfgang Hess who initiated the work on implementing fast speech in unit selection speech synthesis long time ago. Many, many thanks to Prof. Dr. Petra Wagner for taking over supervision when it was about time. She not only lent me her ear, but more importantly also lent me her voice. Thanks for being a great supervisor and a good friend for all this years! And thank you also to my second supervisor, Prof. Dr. Bernd Möbius, for never loosing his patience.

To those who encouraged me to follow my dreams after all, to my long-term companions, friends, supporters and colleagues, at university, in speech technology and in real life - Stefan, Christine, Beate, Bernhard, Nina, Kurt, Hyeonyoung, David, Sebastian, Denis, Charlotte, Katie und Andreas: Thank you for your encouragement, your advice, support and company during all those years!

Danke auch an meine ehrenamtlichen Kolleginnen und Kollegen vom Deutschen Roten Kreuz für ihre Aufgeschlossenheit meinen Hörexperimenten gegenüber, sowie den Umfrageteilnehmern und Testhörern vom Berufsförderungswerk Düren gGmbH, der Rheinischen Schule für Blinde (LVR-Louis-Braille-Schule) Düren, der blista Blindenstudienanstalt - Kompetenzzentrum für Menschen mit Blindheit und Sehbehinderung Marburg, und den Schülerinnen und Schülern der Carl-Strehl-Schule (CSS) der blista Marburg, ohne deren Mitwirkung die Erstellung dieser Arbeit nicht möglich gewesen wäre.

Danke meiner ganzen Familie für die vorbehaltlose Unterstützung. Besonderer Dank gilt meiner Mutter, die immer an mich geglaubt hat und immer für mich da war.

Timo, es gibt keinen Weg, der nicht irgendwann nach Hause führt.

Besiegt ist nur, wer den Mut verliert.
Sieger ist jeder, der weiter kämpfen will.

Franz von Sales

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech synthesis is part of the everyday life of many people with severe visual disabilities. For those who are reliant on assistive speech technology the possibility to choose a fast speaking rate is reported to be essential [Granström, 1991], [Fellbaum, 1996], [Portele and Krämer, 1996], [He and Gupta, 2001], [Asakawa et al., 2002], [Nishimoto et al., 2006], [Borodin et al., 2010], [Ahmed et al., 2012], [McCarthy et al., 2013]. But also expressive speech synthesis and other spoken language interfaces may require an integration of fast speech [Chalamandaris et al., 2010]. Architectures like formant or diphone synthesis are able to produce synthetic speech at fast speech rates, but the generated speech does not sound very natural. Unit selection synthesis systems, however, are capable of delivering more natural output. Nevertheless, fast speech has not been adequately implemented into such systems to date. Thus, the goal of the work presented here was to determine an optimal strategy for modeling fast speech in unit selection speech synthesis to provide potential users with a more natural sounding alternative for fast speech output.

To model fast speech in unit selection speech synthesis the characteristics of natural fast speech production have to be examined at first. The phonetic characteristics of natural fast speech differ from those of speech produced at normal speech rates. The faster somebody speaks the less intelligible their utterances become. Therefore, the characteristics of natural fast speech production are detailed in chapter 2. The term speaking rate is discussed in contrast to similar concepts. Different approaches to measure speaking rate are outlined. Afterwards, the manifestation of changes in speaking rate as well as their effects on different linguistic units are described in section 2.2. The articulatory and acoustic characteristics of natural fast speech are explained, and alterations in characteristics of single segments (section 2.2.2) as well as of larger linguistic units such as syllables, words and phrases are discussed (section 2.2.3). In the last section 2.3 of the first chapter, several strategies to produce different speaking styles are described. The speaking style required for the research to be conducted here is defined as fast and clear at the same time. Speakers are able to produce this speaking style by enhancing the articulatory effort while increasing speech tempo. That way, undesirable phenomena like coarticulation and reduction, known to occur in fast speech, can be avoided as much as possible.

The next chapter 3 of the work presented here deals with speech synthesis itself. In section 3.1, different synthesis techniques will be examined. Parametric synthesis is opposed to data-driven approaches (sections 3.1.1 and 3.1.2). The speech synthesis system applied in the current research, the Bonn Open Speech Synthesis (BOSS), will be outlined in more detail in chapter 3.1.2. Subsequently, different approaches of modeling speaking rate in speech synthesis will be discussed in section sec:modelingspeakingrate. An adequate duration prediction enhances the perceived naturalness of synthetic speech [Carlson et al., 1979], [Brinckmann and Trouvain, 2003]. However, the duration of speech segments is affected by many different factors to be taken into account. Several models have been developed to describe and predict the duration of speech units by considering those factors to different extents (cf. section 3.2.1). The most common models will be discussed and the method of duration prediction to be deployed in the current project will be defined. Afterwards, algorithms to accelerate (synthetic) speech to higher speaking rates, even higher than natural ones (henceforth ”ultra-fast”), are detailed in section 3.2.2. One common option to model fast speech in speech synthesis is to accelerate the speech generated at normal speaking rate by means of linear duration manipulation as described in section 3.2.2. The produced output often shows artifacts known to appear when applying algorithms such as TD-PSOLA [Moulines and Charpentier, 1990], [Liu and Zeng, 2006], and does not sound very natural. Nevertheless, these algorithms are robust and require little computing effort, and are therefore widespread and commonly used in speech synthesis. Thus, they will also be applied to generate the fast and ultra-fast stimuli for the research outlined here.

The perception of fast and/or synthetic speech needs to be taken into account as well when implementing this speaking style in unit selection speech synthesis. Therefore, chapter 4 describes the most important aspects of speech perception. At first, the perception of natural fast speech is detailed in section 4.1. Common models of speech perception as well as investigations about units of speaking rate perception are discussed. Then, section 4.2.1 presents different methods to evaluate artificially produced speech in general. Afterwards, the perception of time-compressed natural speech is considered in 4.2.2, before the perception of synthesized (fast) speech is outlined in section 4.2.3. To conclude with, in section 4.3 it will be outlined how the perception of speech - and in particular the perception of speaking rate and synthesized speech - is influenced by individual listener characteristics [Möller, 2000], [Jekosch, 2005], [Black and Tokuda, 2005], [Syrdal et al., 2012]. Since familiarity with the presented material is one of the listener characteristics to be taken into account, also the judgments collected in the current research will be differentiated by listener groups and synthesis application: the results of sighted or novice listeners are opposed to judgments by trained, mostly visually impaired or blind users for stimuli generated by means of different speech synthesis systems and inventories (cf. chapter 8.2.1).

When preparing the empirical studies to be conducted for the research presented here some fundamental questions arose: What do the blind and visually impaired seek for with respect to the quality of synthetic speech? Do they prefer a monotonous, fast speech synthesis that is prosodically relatively close to natural fast speech as suggested by [Fellbaum, 1996]? Do they not mind a lack in naturalness as long as acoustic transitions important for segment identification are adequately modeled, as in formant synthesis [Moos and Trouvain, 2007]? What kind of speech quality do they prefer in general [Granström, 1991], [Brinckmann and Trouvain, 2003], [Stent et al., 2011]? An early study about speech synthesis applications for the blind conducted by [Portele and Krämer, 1996] revealed that intelligibility was the crucial factor in judging the quality of synthesized speech. However, the researchers noted that the generated utterances often lacked the naturalness of human speech. Nonetheless, for many subjects the ease of use, flexibility, and robustness of a system were at least as important as speech quality in terms of naturalness. In more recent studies, [Chalamandaris et al., 2010] and [McCarthy et al., 2013] confirmed these findings. [McCarthy et al., 2013] observed that for novice users the main drivers of adoption of a certain screen reader software were a human sounding voice as well as the voice quality in general, whereas the most important factors for advanced users were application support and the possibility to speed up the uttered speech to a certain extent. Moreover, advanced users were more comfortable with non-human-sounding speech than novice users. Also [Chalamandaris et al., 2010] who presented a unit selection text-to-speech (henceforth "TTS") synthesis system optimized for use in screen readers in Greek stated that TTS technology in general needed options for adaptation and customization for dedicated applications. Moreover, their interviewees suggested to provide an option allowing for degraded speech quality in exchange for increased speed.

Since the preferences of blind and visually impaired users concerning speaking rate and naturalness of synthesized speech had not been investigated as explicitly as it would have been desirable in order to design an optimal strategy for modeling fast speech in unit selection speech synthesis (cf. also [Stent et al., 2011]), as well as to avoid the problem of a "lack of understanding the users' needs" [Wagner, 2013] a preliminary survey was performed at the beginning of the research presented here [Moers et al., 2007]. Its goal was to find out more details about the requirements and expectations of people using German speech synthesis as assistive technology on a daily basis before starting the main work on integrating fast speech in a unit selection speech synthesis system. Issues and results of this study will be outlined in chapter 5.

The quality of the speech produced by a unit selection speech synthesis system is mainly determined by the inventory speaker [Syrdal et al., 1997].

Therefore, requirements for the selection of a suitable speaker to record a fast speech corpus are derived from the specific speaking style described and defined previously in chapter 2.3. A speaker will be searched for who is able to produce the required speaking style "fast and clear" in an optimal way. The procedure of speaker selection and evaluation will be outlined in chapter 6. After selecting a suitable speaker, the acoustic characteristics of fast as well as fast and clear speech produced by the selected speaker will be investigated and compared to the characteristics of the speaker's normal rate speech. In section 6.2.2, a perceptual evaluation of the different fast speaking styles will be described and results will be discussed (cf. [Moers and Wagner, 2008], [Moers and Wagner, 2009]).

In order to investigate the modeling of fast (and clear) speech in unit selection synthesis two independent but, in terms of linguistic content, identical unit selection inventories will then be created: one in normal and one in fast and clear speech rate. The procedure of corpus recordings is outlined in section 7.1. Further development steps of the two parallel speech corpora are described in chapter 7. In order to build a useful and manageable inventory for fast speech, it will then be investigated whether fast speech utterances articulated as accurately and clearly as possible have a perceptual disadvantage compared to accelerated normal speech rate utterances by contrasting them in a perceptual evaluation (section 7.1.1, cf. [Moers et al., 2010c], [Moers et al., 2010a]). Implications for the implementation of a fast and clear speech corpus into a unit selection speech synthesis system will be discussed subsequently. Thus, the aim of the work presented here was also to integrate the insights of [Lindblom, 1990] and a more flexible approach to inventory creation for unit selection synthesis in order to achieve synthetic speech that is both maximally natural and maximally fast.

The preparation of the unit selection inventory is one of the most time consuming steps during the development of new corpora for unit selection synthesis, as usually a lot of manual labeling is required. To label speech in normal speaking rate automatic labeling techniques are preferred. Since the quality of the synthesized speech depends on the label timing accuracy (LTA) as well [Kominek et al., 2003], using the same segmentation algorithm for both normal and fast and clear speech utterances might result in a considerably increased amount of incorrect labels for fast and clear speech. If so, automatic segmentation would not be applicable to the fast and clear speech utterances although they were articulated as accurately as possible. Thus, the implementation of fast and clear speech into a unit selection speech synthesis system would not be applicable at all for practical reasons (cf. [Moers et al., 2010c], [Moers et al., 2010a], [Wagner, 2013]). In chapter 7.2, the preparation of the fast speech unit selection inventory is outlined. At first, in section 7.2.1 the applicability of automatic labeling techniques to both normal and fast and clear speech will be examined in detail. The actual processing steps include the adaptation of existing transcriptions to the needs of the BOSS system, the automatic segmentation of the corpus recordings of normal and fast and clear speech into speech units by means of an HTK-based aligner adapted to German [Dragon, 2005], and the subsequent analysis of the label timing accuracy for both corpora. Robust guidelines for integrating a fast speech corpus into a unit selection synthesis system are expected to result from the approach discussed here. With regard to duration prediction, the application of Classification And Regression Trees (CART, [Breiman et al., 1984]) to create segment duration prediction models for the normal and the fast and clear speech corpus separately is outlined. Taking into account important phonetic and prosodic features influencing segmental duration, results of a comparative analysis of the generated CART-based duration prediction models for both corpora will be presented in chapter 7.2.2. Conclusions will be drawn whether an adaptation of the duration prediction module to fast and clear speech is required when implementing this speaking style in unit selection speech synthesis.

Blind and visually impaired users of screen reading software seem to prefer the less natural sounding formant synthesis over the more natural sounding unit selection synthesis across all speaking rates (cf. chapter 5, [Moos and Trouvain, 2007]). Next to pure habituation due to repeated exposure [Jannedy et al., 2010], the unproblematic replication of fast and smooth transitions in formant synthesis as opposed to unit concatenation may play a vital role in this preference. As [Winters and Pisoni, 2004] pointed out, the advantage of formant synthesis might disappear when concatenative synthesis with larger units is used. Therefore, the next step in the work reported here is the definition of the adequate unit size to synthesize fast speech. The approach suggested by [Breuer and Abresch, 2004] to treat phone sequences which are prone to heavy coarticulation as atomic in the sense that they are regarded as two or more phones, but one indivisible synthesis unit will be taken up in the investigation outlined in chapter 8.1. It is expected to find a possible solution for modeling fast speech both more naturally - by using prerecorded concatenation units - and more intelligibly by including typical smooth transitions in heavily coarticulated contexts in order to achieve synthetic speech that was both maximally natural and maximally fast.

Finally, after defining the adequate unit size to synthesize fast and clear speech, the intelligibility, naturalness, and overall acceptability of utterances generated from the different underlying corpora and with different synthesis systems at different speaking rates will be evaluated. To investigate their characteristics, Semantically Unpredictable Sentences (SUS, [Benoit and Grice, 1996]) will be used generated with both unit selection inventories - normal and fast and clear speech - as well as with formant synthesis (cf. [Syrdal et al., 2012]). Afterwards, a Mean Opinion Score (MOS) will be collected from two different listener groups: trained blind and visually impaired daily users of screenreader software, and untrained, mostly sighted listeners as a control group to accommodate the possible bias for the trained blind and visually impaired users regarding formant synthesis. Additionally, the Word

Error Rate (WER) will be analyzed depending on listener group, synthesis system and speaking rate. The results of the perceptual evaluation of the speech synthesized from different underlying corpora and different systems at different speaking rates will be outlined and discussed in section 8.2.

All the mentioned aspects of implementing fast speech as a separate speaking style in unit selection speech synthesis show that there are several open questions about the adequate treatment of different speaking styles, especially fast speech, and their applicability in concatenative speech synthesis systems. Taking the outlined requirements and prerequisites into consideration, the research presented here aims at defining robust guidelines for integrating fast and clear speech as a separate speaking style into a unit selection synthesis system. At the same time, the generated speech shall be suitable and acceptable for visually impaired users of such assistive speech technology in future.

Parts of this thesis and the work presented therein have been published in the following articles:

- Moers, D., Wagner, P., and Breuer, S. (2007). Assessing the adequate treatment of fast speech in unit selection systems for the visually impaired. Proceedings 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-6), Bonn, Germany.

- Moers, D., and Wagner, P. (2008). Evaluation eines Sprechers fÃ¼r schnell gesprochene Sprache in der Unit-Selection basierten Sprachsynthese. ITG-Fachtagung Sprachkommunikation, Aachen, Germany.

- Moers, D., and Wagner, P. (2009). Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis. Proceedings Interspeech 2009, Brighton, UK.

- Moers, D., Wagner, P., Möbius, B., Müllers, F. and Jauk, I. (2010). Integrating a fast speech corpus in unit selection speech synthesis: Experiments on perception, segmentation and duration prediction. Proceedings Speech Prosody 2010, Chicago, IL, USA.

- Moers, D., Wagner, P., and Möbius, B. (2010). Erzeugung schnell gesprochener Sprache in der Unit-Selection-Sprachsynthese. Proceedings ESSV 2010, Berlin, Germany.

- Moers, D., Wagner, P., Möbius, B., and Jauk, I. (2010). Synthesizing Fast Speech by Implementing Multi-Phone Units in Unit Selection Speech Synthesis. Proceedings 7th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-7), Kyoto, Japan.

- Moers, D., Wagner, P., Möbius, B., Müllers, F. and Jauk, I. (2010). Schnell gesprochene Sprache in der Unit-Selection-Sprachsynthese: Untersuchungen zu Korpuserstellung und -aufbereitung. ITG-Fachtagung Sprachkommunikation, Bochum, Germany.

- Moers, D. (2011). Schnell gesprochene Sprache als Einheiten-Auswahl-Inventar in der Unit-Selection-Sprachsynthese. Proceedings ESSV 2011, Aachen, Germany.

Chapter 2

Fast Speech Production

When investigating the modeling of fast speech in unit selection speech synthesis, the characteristics of natural fast speech have to be considered. At the beginning of this chapter in section 2.1 the term *speaking rate* is outlined and contrasted with similar concepts. Different approaches to measure speaking rate are discussed. The manifestation of changes in speaking rate as well as the articulatory and acoustic characteristics of natural fast speech are described in detail in section 2.2. Alterations in characteristics of single segments (section 2.2.2) as well as of larger linguistic units such as syllables, words, and phrases are discussed subsequently (section 2.2.3). In section 2.3, several strategies to produce different speaking styles are illustrated. Requirements for the suitability of a speaker to record a fast speech corpus are derived from a specific speaking strategy. The selection procedure itself is presented in detail later on in chapter 6.

2.1 Speaking Rate Definition and Quantification

The *speaking rate* is defined as the number of linguistic units produced per time unit including pauses. It is measured over the course of a stretch of speech. In contrast, the term *articulation rate* refers to the number of linguistic units produced per time unit excluding pauses. Thus, the articulation rate is measured over the course of an interpause stretch, an intonation phrase, or an entire utterance. It can be seen as one constituent of the speaking rate. The other constituent of the speaking rate is the *pause rate*, which describes the number and frequency of their occurrence [Miller et al., 1984]. [Laver, 1994] modified this definition to include non-linguistic material like hesitations and filled pauses into the articulation rate. Only silent pauses were not taken into account. According to his definition, speaking rate refers to the tempo of an entire utterance including all linguistic and non-linguistic material. [Fant et al., 1992], however, made another distinction: According to them, *speech tempo* is a relational measure reflecting a concrete, specific speaking rate compared to a reference speaking rate whereas the ”speech rate” is measurable in terms of linguistic units per stretch of speech.

In early investigations of speaking rate, its changes were mainly attributed to a modification of the number and duration of pauses contained in an utterance. [Goldman-Eisler, 1968], for example, distinguished between the ”rate of talking” being a measure of the tempo of planning an utterance, and the ”rate of articulation” where pauses, hesitations, and tongue slips were excluded from measurement. The time needed to produce speech, however, was considered constant. That would mean that a faster speaking rate can only be achieved by a decrease of the duration or even a total deletion of pauses. In a follow-up study, [Grosjean and Deschamps, 1975] confirmed the assumptions of [Goldman-Eisler, 1968] at first, but after analyzing the same underlying data again [Miller et al., 1984] argued that the turns of speech used by [Goldman-Eisler, 1968] to measure the rate of articulation were way too long. Instead of measuring the rate of articulation over a group of 30 syllables the authors calculated it over shorter paragraphs of speech between pauses. They observed that the rate of articulation varied to a great extent between speakers as well as between utterances of a single speaker. Similar observations were made by [Adank and Janse, 2009] who noted that in conversation, speakers varied their speaking rate between 140 and 180 words per minute. Eventually, [Miller et al., 1984] concluded that a change in ”rate of talking” had to be ascribed to both a change in rate of articulation and a change in rate of pauses.

Also [Crystal and House, 1990] observed a huge variability of articulation rate measured as the average syllable duration for interpause intervals which they called ”runs”. Here, the articulation rate was neither random nor talker-idiosyncratic but dependent on the content of the run, and at the same time a function of the syllabic and stress characteristics of the specific materials. In another investigation, [Trouvain and Grice, 1999] found in addition that the speaking rate showed larger differences between speech samples comprising the same content and produced at slow, normal, and fast speech tempos than the corresponding articulation rate. The most extreme speaker accelerated the speaking rate for fast speech by 28% whereas the articulation rate was only 18% faster than normal speech. The researchers concluded that the changes in speaking rate were attributable to changes in pause duration to a large extent. In line with those findings, [Trouvain et al., 2001] showed that the most precise way to determine the articulation rate in spontaneous speech was to measure it over stretches of speech not containing any pauses: The authors found the highest correlation between the number of linguistic units and time needed to articulate those for so called inter-pause stretches. The correlation was much smaller for intonation phrases which optionally included pauses. As intonation phrases are sometimes longer and sometimes shorter than inter-pause stretches, the result was not attributable to a generally longer duration of the intonation phrases.

Explanations by [Wood, 1973a] imply that a discrimination between *gross* and *net* measures of speaking rate has to be made where ”net measure of rate” refers to the units actually produced in an utterance, and ”gross measure of rate” describes the number of units produced including pauses. Additionally, the author stated that there might be definitions where a difference is made between speaking rates counting either units being actually represented in the speech signal investigated (”concrete”), or units derived from some idealized underlying construction (”abstract”). Those kind of definitions can also be found in the investigations conducted by [Trouvain et al., 2001], [Engstrand and Krull, 2001], [Dellwo and Wagner, 2003] and [Koreman, 2006]. Here, the different measures of speaking rate are called *intended rate* (ISR) and *realized* or *laboratory measured speech rate* (LSR). For [Engstrand and Krull, 2001], the articulatory reduction process is reflected in the distinction between ”underlying syllables” and ”phonetic syllables” where the latter correspond to realized syllables. [Koreman, 2006] additionally defines an *Articulatory Precision Index* (API) which reflects the relative deletion rate of linguistic units for fast versus slow speakers. [1]

[Kohler et al., 1981] assumed that the temporal organization of speaking took place on several hierarchical layers. According to them, speaking rate belonged to the ”macro layer” of an utterance. The ”micro layers”, consisting of syllable chains and phonemes, had to be integrated into this macro layer during speech production. Moreover, they distinguished between ”Globaltempo” (global tempo) reflecting the average speaking rate of an utterance, and ”Momentantempo” (instantaneous tempo), the local speech tempo of the actual realization. This approach was adopted and extended by [Pfitzinger, 1996], [Pfitzinger, 1998]: He distinguished between *global*, *local* and *relative* speaking rates. In his model, the ”global speaking rate” provides information about the average speaking rate of an utterance. It reflects the proportion of the total number of linguistic units to the accumulated total duration of the linguistic units. This is consistent with the definition of speaking rate by [Laver, 1994]. The ”relative speaking rate” describes the proportion of individual speaking rates of the same utterance to each other. It was originally defined by [Ohno and Fujisaki, 1995], [Ohno et al., 1997] who tried to avoid the problem of exact determination of segment boundaries when estimating speaking rate in automatic speech recognition applications. The ”local speaking rate” is determined in regular intervals for small sections of an utterance. Also here, the number of linguistic units per time unit is calculated. For a long utterance, the result is a number of local speaking rate measurements which can be represented in the form of a local speaking rate curve. This curve is synchronized with the speech signal, showing a low value for slow parts of the utterance and a high value for fast parts. The most important conclusion drawn by [Pfitzinger, 1996] and [Pfitzinger, 1998] was that the local speaking rate was best reflected by a linear combination of syllable rate and phone rate.

In addition to the above mentioned discussions about the definition of

[1] *Fast* and *slow* here refer to the "habitual speaking rate"as defined in [Tsao and Weismer, 1997].

Speaking rate	**words per minute (w.p.m.)**
Fast	above 220 w.p.m.
Moderately fast	190 to 220 w.p.m.
Average	160 to 190 w.p.m.
Moderately slow	130 to 160 w.p.m.
Slow	below 130 w.p.m.

Table 2.1: Categorization of speaking rate in words per minute after [Pimsleur et al., 1977], cited after [Tauroza and Allison, 1990].

speaking rate and the stretch of speech to refer to for its measurement, there is also disagreement about the optimal linguistic unit of measurement. In their investigation of the role of speaking rate in foreign language teaching and learning, [Tauroza and Allison, 1990] first reverted to measuring speaking rate in words per minute and defined the average speaking rates for certain speaking rate categories as listed in table 2.1.

After analyzing several different speaking styles and having discovered much variation in mean word lengths between different speaking styles, [Tauroza and Allison, 1990] put up for discussion which unit of measurement would be most suitable to determine the speaking rate as words per minute did not seem to be the best choice. Comparing words per minute to syllables per minute as a rate measure, the authors observed that the average word length was notably shorter than reported by [Pimsleur et al., 1977] in three of the four speaking style categories. Therefore, they finally opted for *syllables per minute* as the best measure to determine speaking rate when comparing different speaking styles. This approach is similar to the one followed earlier by [Grosjean and Deschamps, 1972] who also suggested to determine the speaking rate by referring to syllables as underlying linguistic units. Based on this assumption, the authors calculated the speaking rate, the articulation rate, and the duration of pauses as well as the average duration of utterances, the average duration of pauses, and the proportion of articulation rate to duration of utterances realized by French and English native speakers. They concluded that the observed differences in speaking rates between those two languages were neither attributable to a difference in the length of pauses nor to the rate of articulation itself, but rather to a smaller number of pauses inserted by French speakers compared to English speakers.

[Neppert and Petursson, 1986] referred to the syllable as underlying unit for speaking rate measurement as well and proposed a categorization of speaking rate in terms of syllables per second as depicted in table 2.2. In contrast, [Fant et al., 1992] stated that the average syllable duration - which according to them equals the inverse of syllables per second - was not a reliable measure because of the different complexity of words. Instead, the authors considered measuring speaking rate based on the average phone duration. Local variation of speaking rate was only conceded to a small degree, mostly related to emphasized content

Speaking rate	**syllables/second**
very slow	2,9 to 3,0
decreased/slow	3,1 to 3,5
normal	4,5
increased	5,0
fast/very fast	5,6 to 6,0

Table 2.2: Categorization of speaking rate in syllables per second after [Neppert and Petursson, 1986].

words and final lengthening.

The research performed by [Faust, 1997] was based on the assumption that the preferential linguistic unit for speaking rate determination was indeed the phone. Similarly to [Tauroza and Allison, 1990] and [Grosjean and Deschamps, 1972] who were arguing for the preference of syllable length over word length when measuring speaking rate, [Faust, 1997] stated that syllables often were reduced and therefore showed too much variability in duration and complexity to serve as a useful measure of speaking rate. Investigations conducted by [Roach, 1998] followed this argumentation and confirmed that syllables were very different in their complexity and therefore not the optimal unit to measure speaking rate. In a later study, [Trouvain et al., 2001] investigated the correlation between duration and number of segments for different stretches of speech. As was expected, the highest correlation was found for the smallest linguistic units under investigation: realized phones. In contrast, the largest linguistic unit examined, the word, showed only a small correlation between duration and number of segments. The authors ascribe this finding to the higher duration variability of words. However, the approach to refer to the number of phones per time unit as speaking rate determiner was not new. It can also be found in much older investigations, for example in [von Essen, 1949].

The preceding explanations show that there is a consensus in the literature on determining speaking rate by distinguishing between stretches of speech including or excluding pauses, but in other details concepts differ significantly. Especially the underlying linguistic unit to be referred to for measurements are a matter of ongoing discussions. For the work presented here, the speaking rate will be determined in underlying syllables per second for an entire utterance as with regard to fast speech the determination of speaking rate in (realized) phones per second could be difficult and misleading [Trouvain et al., 2001]. In addition, since a certain variation is to be expected within a bunch of utterances produced at the same intended speaking rate [van Santen, 1992], [Pfitzinger, 1998], [Wang et al., 2000], the estimation of the speaking rate in syllables per second over a complete utterance seems to be a more adequate measure to demonstrate the global characteristics of the fast speech produced here. The speaking rate measured will be categorized in accordance with the

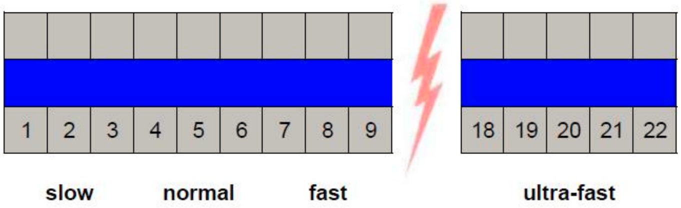


Figure 2.1: Categorization of speaking rate in syllables per second after [Moos and Trouvain, 2007].

definition provided by [Moos and Trouvain, 2007] and depicted in 2.1, to be able to also define speaking rates beyond naturally producable speaking rates.

2.2 Natural Fast Speech Production

To investigate the modeling of fast speech in unit selection speech synthesis, the manifestation of changes in speaking rate and their impact on different linguistic units need to be examined first. It can be assumed that fast speech differs significantly from speech produced at a normal speech tempo both in articulatory and acoustic characteristics. In fast speech, the articulation of segments has to take place in a smaller time frame than in speech uttered at a normal speech tempo. Linguistic units are produced with more gestural overlap and acoustic interference. Therefore, the phenomena of *coarticulation* and *reduction* are recalled in section 2.2.1. Because of their increased occurrence in accelerated speech, the quality and quantity of vowels as well as of consonants changes dramatically. Additionally, the transitions between single speech units are altered. Those manifestations of speaking rate changes are described in the subsections of chapter 2.2.2. But also larger prosodic units like syllables or words are influenced by a change in speaking rate. Prosodic features like intonation, phrasing and pausing are affected as well, as explained in chapter 2.2.3. It will be shown in section 2.3 that it is highly context and speaker dependent whether or not the above mentioned phenomena appear in fast speech.

2.2.1 Coarticulation and Reduction

Speech signals do not consist of sequences of separate speech units simply concatenated one after the other, but rather of an ongoing speech flow of conglomerated phonemes. Phenomena like coarticulation, assimilation, and reduction are likely to occur, even at a normal speaking rate. They make it almost impossible to cut speech signals into distinct single units.

Coarticulation can appear in both directions on the time axis. ”Progressive” or perseverative coarticulation occurs where a segment is influenced by the preceding unit. If the realization of a phoneme is influenced by the following speech unit, it is called ”regressive” or anticipatory coarticulation. According to [Neppert and Petursson, 1986], progressive coarticulation usually is less incisive than regressive coarticulation. Additionally, progressive coarticulation is more passive, whereas anticipatory coarticulation has its source in speech motor planning [Neppert and Petursson, 1986].

A slightly different definition of coarticulation can be found in [Daniloff and Hammarberg, 1973]: The authors describe progressive coarticulation as ”carry-over coarticulation” only related to physiological factors whereas anticipatory coarticulation arises from an active planning process which has access to an idealized phoneme representation on a higher linguistic level. This idealized phoneme form defines articulatory targets through several parameters. The active planning process is called ”accommodation”. There are two types of accommodation to distinguish: One is the adaptation of the place of articulation between neighboring segments which are articulated with the same articulator (*assimilation*), and the other one is characterized by the observation that for neighboring segments articulated with different articulators the free articulator either already moves into the direction of the upcoming phoneme, or still remains in the position of the preceding one.

[Whalen, 1990], however, discussed whether coarticulation is actually part of the motor program or rather a consequence of executing it. He asked if modifications of coarticulation were due to feature-spreading or to temporal concurrence, to the overlap of prominence curves or to the phasing of gestures. During his investigations, anticipatory coarticulatory effects did not appear when only the beginning of a predefined utterance was known whereas perseverative coarticulatory effects were observable under both conditions, known and unknown continuation of an utterance. The researcher concluded that coarticulation must either be of mechanical origin or can be planned on short term during the articulation of an utterance onset. As a consequence of his observations, he stated that

> coarticulation, though presumed to be due to the constraints of producing speech in real time, is largely a result of planning an utterance rather than an automatic consequence of successfully producing that utterance. [Whalen, 1990].

Together with coarticulation, the *reduction* of speech units is a common phenomenon to observe in natural speech production as well. For vowels - which are more affected by reduction than consonants - the term ”reduction” most of the time refers to target frequencies of a generic production of the respective vowel which are not reached anymore (”target undershoot”) [Lindblom, 1963], [Greisbach, 1991], [Greisbach, 1992], [Kohler, 1995]. It is a matter of ongoing discussions whether the noticeable shift of the formant frequencies is a movement directed towards the center of the vowel space (”centralization”), or simply a consequence of mutual influence between neighboring segments (”coarticulation”) [Lindblom, 1963], [van Bergem, 1993], [Kohler, 1990], [Aylett, 2000], [Weiss, 2008]. [van Bergem, 1995] interpreted vowel reduction as a tendency towards ease of articulation by contextual assimilation. In that, he distinguished between ”lexical reduction” and ”acoustic reduction” where the latter reflects the expected decrease in formant frequencies. According to him, the central vowel [ə] is not produced with a neutral vocal tract but very much dependent on the surrounding consonants, and therefore to be seen as the most economical direct movement from the preceding consonant to the following consonant (cf. also [Barry, 1998]). [van Bergem, 1995] concluded that also full vowels are not reduced to some neutral vowel in the center of the vowel space, but to an articulatory position that is dependent on the consonantal context. In accordance with [Gopal, 1990] and [Lindblom, 1963], [Barry, 1998] diagnosed a ”context effect” especially for [@] . However, in fast speech he found a dominance of the vocalic context factor as opposed to the dominance of the consonantal factor in the slow condition he investigated. The author deduced an increasing strength of different flanking vowels with increasing tempo and stated that ”the schwa is a vowel without articulatory target that is completely assimilated with its phonemic context.” [Barry, 1998]. Additionally, he noticed that the first formant F1 generally showed lower values in fast speech attributable to less mouth opening which is in contrast to [Koopmans-Van Beinum, 1990] who found higher F1 values indicating a more open articulation (cf. section 2.2.2).

Consonants are influenced by an acceleration of speaking rate, too. Like vowels, they are often shortened in duration. However, due to the fact that certain consonants are already quite short, for example plosives, and therefore cannot be compressed more without losing the segment’s main characteristics, consonantal shortening and reduction is much less pronounced [Gay, 1981], [Pasdeloup et al., 2008] compared to vowel reduction. Still, different types of consonants are affected in different ways by speech rate acceleration. Plosives, for example, become weaker which means that the characteristic closure is not completely realized anymore resulting in a lack of pressure which in turn leads to plosive bursts performed with less intensity. Consequently, in fast speech the acoustic characteristics of plosives are getting more similar to those of approximants [Kohler, 1990]. Nevertheless, [Greisbach, 1991] and [Caspers and van Heuven, 1992] noted that there are necessary elements of consonant articulation that have to remain intact also in fast speech to reach the communicative goal, like for example word-initial plosives. In general, consonant reduction is less likely to occur than vowel reduction, and findings about consonant reduction often are only a side product of broader studies designed to investigate vowel reduction.

2.2.2 Articulatory and Acoustic Effects

In addition to shortening of speech units in terms of duration an extended articulatory overlap can be observed in fast speech [Byrd and Tan, 1996]. The increasing overlap of articulatory gestures as well as the limited movement velocity of the articulators are sources of intensified assimilation and reduction when speaking rate increases. [Engstrand, 1988] and [Engstrand and Krull, 2001] reported that they noticed "quite drastic" reduction phenomena in fast speech: increased consonant deletion, vowel merging, vowel elision, palatalization (cf. [Cooper et al., 1983]), and nasalization. All these phenomena affect the quality of the speech produced, and consequently have an impact on the listeners' perception. Therefore, in the following section articulatory changes occurring when fast speech is produced are described, followed by a discussion of the acoustic manifestations of those changes. Their impact on listeners' perception will be discussed later on in chapter 4.1.

Changes in Articulation

Already [Gay, 1981] noticed changes in articulatory displacement, articulatory velocity, and intrasyllablic coarticulation during his investigations of fast speech. According to him the observed phenomena were induced by a reordering of speech motor strategies during fast speech production. In a later study, [Davidson, 2006] confirmed this assumption of a reorganization of the timing of articulatory gestures in fast speech which led to more articulatory overlap or even the disappearance of certain gestures. [Kohler, 1990], however, stated that deletion and assimilation of phonemes were the result of a combination of reorganizing of articulatory gestures and reducing the effort in fast speech production ("criterion of motor economy") on the one hand, and "listener orientation" (perceptual and social constraints, depending on the communicative situation) on the other hand. These hypotheses will be discussed in more detail in section 2.3, based on [Lindblom, 1990]'s theory of hyperspeech and hypospeech production. Generally speaking, in accordance with [Mefferd and Green, 2010], articulatory movements often are underspecified in fast speech which usually is referred to as (articulatory) *target undershoot*.

In contrast to other investigations (for example [Miller and Baer, 1983], [Arons, 1992]), [van Son and Pols, 1989] noticed that for their speaker a higher reading rate resulted in shorter vowel durations and overall higher F1 values, but not in a shrinkage of the *vowel space*. The authors concluded that coarticulation and reduction were not the result of physiological articulatory limitations but language governed features. They deducted this assumption from their observation that the correlation between formant frequency and duration was minimal, and differences for fast spoken vowels compared to normal vowels were negligible. [Jannedy et al., 2010] observed a similar phenomenon: Although they were generally reduced in fast speech, movement amplitudes remained still very large for their subject, a highly trained speaker. From their production experiments, the researchers concluded that articulatory reorganization as well as speech errors were avoided by means of training of repeated patterns. This finding will play an important role during corpus recordings, as described in chapter 7.1.

Looking at the single articulators, it becomes obvious that the larger and heavier the articulator, the smaller its velocity and mobility. [Fuchs and Perrier, 2005], for example, found that jaw oscillations were reduced and the jaw was held in a relatively high position when speaking rate was accelerated. [Ostry and Munhall, 1985] measured the lowering gesture of the tongue for consonant-vowel sequences with alternating speech rate. They found reliable correlations between the amplitude of the tongue dorsum movement and its maximum velocity. Moreover, the ratio of the maximum velocity to the gesture's extent, seen as indicator of the articulator stiffness, varied inversely with the duration of the movement. This relation was stable within and across different conditions. The authors concluded that the speaking rate influenced the maximum-velocity-to-movement-amplitude ratio in a systematic manner. However, they also noticed that changes of speech rate were produced differently by different subjects. [Adams et al., 1993] investigated the velocity profiles of the movements of the lower lip and tongue tip during the production of stop consonants. The researchers found changes in the topology of the speech movement velocity-time function for fast speaking rates. These were associated with changes in motor control strategies: The unitary movements that were observed may have been predominantly pre-programmed. It was also noted that opening gestures showed more consistent changes than closing gestures. The changes in duration of the specific movement were generally different for the lower lip versus the tongue tip.

Due to the physiological constraints explained previously, the maximum articulation rate is limited. For German, for example, [Greisbach, 1992] and [Jannedy et al., 2010] reported a maximum articulation rate of 9-11 syllables per second for trained speakers. [Jannedy et al., 2010] noted in addition that this amount was highly dependent on the consonantal complexity of the syllables involved. For Dutch, [Janse et al., 2003] observed articulation rates of 6.7 syllables per second for normal speech and 10.5 syllables per second for fast speech. In contrast, for syllable timed languages higher articulation rates are documented in the literature. For instance, [Martinez et al., 1997] reported fast articulation rate for Castilian to be at 12-15 phones per second. For French, [Pasdeloup et al., 2008] observed a fast articulation rate of 15.31 phones per second, whereas normal rate included 12.33 phones per second, and slow rate comprised 9.88 phones per second. These rates are similar to what was reported by [Dellwo and Wagner, 2003] for their fast speech recordings in the "Bonn Tempo" corpus. However, [Roodenrys et al., 2002] noted that a higher word frequency, a larger size of the phonological word neighborhood, and a higher neighborhood frequency facilitated the maximal articulation rate. The authors also observed noticeable differences in articulation rate when a real word was produced as opposed to a non-word. Taking these findings together, it becomes obvious that the maximum speaking rate which can be achieved by a human speaker is variable, dependent on certain conditions, but generally speaking only possible within a limited range. Going beyond this natural limit means to investigate ”super human speech rates” ([Granström, 1991]). Such super human speech rates were categorized as ”ultra-fast” by [Moos and Trouvain, 2007]. This naming will also be applied in the current research when it comes to the generation and evaluation of speaking rates beyond natural human fast speech production as described in chapter 8.

Increased coarticulation and reduction in fast speech result in a severe change of the acoustic characteristics of single phones and the transitions between them, as well as of larger prosodic units like syllables, phrases, and pauses. Additionally, pitch movements, fundamental frequency, and overall intonation are influenced. Those effects are described one by one in the following sections.

Acoustic Effects on Segmental Level

When speaking rate increases, vowels are shortened in overall duration. Vowels can roughly be described as to consist of three parts: The ”onset” at the beginning of a vowel which includes the formant movements (”transitions”) from the preceding sound, followed by the biggest section in the middle of the vowel where the formant frequencies stay almost stable, optimally after having reached commonly defined target frequencies (often also referred to as ”steady state”), and the ”offset” which includes the transitions to the following sound. Since the transitions at the onset and the offset of a vowel are very important for the vowel's perceptual identification ([Lehiste, 1972], [Martinez et al., 1997]), they may not be curtailed or even left out (cf. section 2.2.2). , the part of a vowel the most affected by shortening is the middle section where the formant frequencies often stay quite constant for a while. [Port, 1981] found that vowels were shortened by 24% in fast speech; [Gay, 1968] and [Gay, 1978] noted similar measures which stayed constant throughout different speaking rates. In contrast, [Martinez et al., 1997] reported that vowels were shortened by 47.5% in normal speech compared to slow speech, and by 61.9% in fast speech compared to slow speech. According to [Lehiste, 1972] (after [Port, 1981]), at some point vowels become *incompressible* as they approach a minimum duration. This effect was described in the duration model developed by [Klatt, 1976], [Klatt, 1979] who hypothesized that a minimum duration was required to execute articulatory movements (cf. chapter 3.2.1). In a later study, the findings of [Windmann et al., 2013] supported and generalized those findings on incompressibility at increased speaking rates.

[Hoole et al., 1994] analyzed different patterns of compression for German tense versus lax vowels over changes in speech rate. In German, the distinction between tense and lax vowels is the same as the one between long and short vowels. [Hoole et al., 1994] found a ”duration compression effect” for both vowel groups; however, it was vastly greater for tense vowels: Lax vowels were shortened by 12.5% whereas tense vowels were contracted by 52.3%. That meant that the duration of tense vowels was much more variable than the duration of lax vowels. [Gopal, 1990] and [Crystal and House, 1990] investigated the proportion of durations of tense vowels to the duration of lax vowels, and found that this proportion was independent from the respective speaking rate. Still, [Gopal, 1990] pointed out that this was only true when the following context was not taken into account. If the following context was taken into account, the proportion showed to be very variable across different speaking rates and vowel pairs (cf. [Lindblom, 1963]). The author concluded that ”(...) the proportion of lax vowel duration to tense vowel duration does not stay constant as a function of rate.” [Gopal, 1990]. In contrast, [Benus and Mady, 2010] found that the phonemic quantity contrast for vowels was salient in their data, and minimally affected by lexical stress or speech rate. They noticed a minor but consistent compression of the quantity contrast (”long/short ratio”) in speech produced at a fast rate.

Another important effect occurring in fast speech is vowel reduction. In early investigations it was assumed that an increase of the articulation rate simply caused a horizontal compression of the spectrogram because each segment was shortened in duration whereas formant frequencies were expected to stay constant. The research conducted by [Wood, 1973b], [Kohler, 1990] and [Widera, 2003] revealed that this assumption was not correct - vowels were severely reduced during the production of fast speech. In contrast, [Gay, 1978], [Engstrand, 1988], [Fourakis, 1991], and [Pols and van Son, 1993] noticed that the spectral characteristics of vowels were not significantly influenced by changes of speaking rate, nor was the overall duration. Also [Benus and Mady, 2010] stated that fast speech did not seem to be realized with more centralized vowels than normal speech. In other investigations, [Lindblom, 1963] and [Kuwabara, 1997] observed that reduction was duration-dependent, and directed towards the center of the vowel space. Additionally, [Widera, 2000] found different reduction levels for different vowels which were mainly influenced by the respective vowel duration. [Hirata and Tsukada, 2004] investigated the question whether long vowels were more likely to reach their target frequencies than short vowels in different speaking rate conditions. Since the researchers found their assumption confirmed, they concluded that in general long vowels resisted coarticulation and reduction more than short vowels. Moreover, the authors observed that long vowels also occupied a more peripheral portion of the vowel space making them more distinct than short vowels.

[Lindblom, 1963] formulated a function to describe the so called ”target undershoot” he observed for vowel realizations in fast speech. The author noted that vowel reduction was strongly related to vowel duration as well as to consonantal context. The amount of reduction was determined by overall vowel duration whereas the context had an influence on the direction of the formant shift. However, no effect was observable if only the average formant frequencies of different vowel instances were analyzed. To sum up, [Lindblom, 1963] defined context-dependent ”locus equations” to reflect the distance between the hypothetical formant frequencies characteristic for the specific consonantal context and the (ideal) target frequencies of the vowel. The result was interpreted as an indication for the amount of coarticulation to expect. The larger the distance, the more target undershoot was anticipated. With this approach, the researcher was able to explain about 50% of formant variation he observed in his data. [Gay, 1978] even numbered the amount of target undershoot observable in fast speech for individual instances of vowels he analyzed more precisely: According to him, changes in F1 were approximately 50 Hz, whereas decreases in F2 and F3 were about 75 Hz. [Fourakis, 1991] investigated the effects of stress and speaking rate on vowel reduction by analyzing the production of stressed or unstressed vowels in normal, slow, and fast speech. He found that the effect of changes in stress was slightly larger than the effect of a change in speech rate. The amount of phonetic vowel reduction was determined by calculating the *Euclidean distance* of the actually produced vowel to a vowel produced with a neutral vocal tract. As a conclusion, the author stated that the size of the vowel space was affected by speaking rate changes in all conditions but no major influence of either condition was observable. However, the vowel space was somewhat smaller in the fast-unstressed condition than in slow-stressed condition; the effect of tempo was slightly stronger than that of stress. In contrast, [Miller and Baer, 1983] reported of a ”shrinkage” of the vowel space observable in fast speech. The approach of analyzing vowel reduction by means of the Euclidean distance of the realized vowel to a virtual neutral vocal tract center is also applied in the analysis of the characteristics of fast and clear speech produced by the selected speaker as described in chapter 6.2.1.

Although [Lindblom, 1963]'s observations regarding the so called *target undershoot* phenomenon point to a strong connection between vowel duration and quality, it must nevertheless be assumed that also other factors have an influence on vowel formant frequencies. As discussed earlier, formant frequencies are strongly related to articulatory gestures. Consequently, target undershoot is also expected to occur in situations where less effort is made to produce speech as, for example, in unstressed syllables. Reduction phenomena are therefore more likely to occur in unstressed syllables. Since stress is strongly correlated with duration, one could assume that it is rather a change in stress than a change in duration which causes vowel reduction. Especially in stress-timed languages like German (cf. section 2.2.3), it is important to keep a sufficient contrast between stressed and unstressed syllables for effective communication [Granström, 1991], [Fant et al., 1991]. [Fant et al., 1992] found that in Swedish a distinct reading mode showed a 22% increase of the duration of stressed syllables and an 11% increase of the duration of unstressed syllables as opposed to normal reading. Comparing fast reading to normal reading, it became obvious that unstressed syllables suffered more from the acceleration of the speech tempo than stressed ones: Unstressed syllables were shortened by 10% whereas stressed syllables were shortened by only 5%. Furthermore, the authors noted that the average number of phonemes per syllable was 2.9

for the stressed reference and 2.3 for the unstressed reference, in accordance with the results of [Kuwabara, 1997]. These results were confirmed by a more recent study by [Janse et al., 2003]. The researchers detected that Dutch speakers reduced unstressed syllables more (up to 68%) than stressed syllables (up to 33%). They attributed this observation to the efforts of the speaker to preserve the more informative parts of speech which resulted in a more pronounced prosodic pattern as the relative duration difference between stressed and unstressed syllables was increased that way. [Janse et al., 2000] noted that in Dutch, unstressed vowels were also reduced more than stressed vowels when speaking rate was increased, irrespective of whether the unstressed vowel was a [@] or a full vowel. [Lindblom, 1990], however, stated that unstressed syllables were reduced because they were produced in a shorter time frame, and not because of decreased articulatory effort as it appears in unstressed syllables. In his studies, he found that target undershoot occurred for stressed and unstressed syllables to the same extent. The scientist concluded that inherent limitations to the articulatory system existed which were independent from the articulatory effort but dependent from the specific duration. For American English, [Crystal and House, 1988] observed that across different speaking rates, the proportion of stressed to unstressed vowel duration was of relative nature and stayed constant. In contrast, [Peterson and Lehiste, 1960] and [Gopal, 1990] stated that unstressed syllables showed a stronger shortening in fast speech than stressed syllables. This would mean that the difference in duration between stressed and unstressed syllables was increased in fast speech (cf. [Delattre, 1966], [Hoequist, 1983]). A later investigation of American English by [Crystal and House, 1990] indicated that the proportion of stressed syllables decreased from nearly 75% in normal speech tempo to less than 50% in fast speech but the relative duration of stressed syllables or stressed vowels in a stress group stayed stable, despite the increasing number of unstressed syllables. Moreover, [Widera and Portele, 1999] found that stressed vowels were reduced less than unstressed ones. If the vowel was reduced more, syllables were perceived as less prominent.

It is well known that consonantal shortening and reduction are much less pronounced in fast speech than vowel reduction [Gay, 1981], [Okadome et al., 1999], [Pasdeloup et al., 2008]. Different types of consonants are affected differently by speech rate acceleration. ”Creaky voice”, for example, is used much more often instead of a full glottal stop, especially when two vowels are coming together [Okadome et al., 1999]. The acoustic characteristics of plosives are similar to those of approximants in fast speech [Kohler, 1990]. This means that consonants in fast speech often change their manner of articulation, accompanied by a change of the acoustic parameters. [Byrd and Tan, 1996] observed that all consonants were shortened in fast speech, but to a very different degree. Only the [d] in syllable-initial position was not shortened at all. Shortening and reduction were most obvious if the consonant was part of the coda of a syllable. The researchers noticed that plosives were reduced more than fricatives, and apical consonants were reduced more than velar ones. However, consonant shortening was also very much speaker dependent. Taking a closer look at plosives, it becomes obvious that these segments are affected the most when speaking rate is increased. [Kohler et al., 1981] found that in fast speech, the closure of plosives often was realized only partially or not at all. The homorganic voiced approximant or voiceless fricative were produced instead. The researchers assumed that in fast speech speakers reach their articulatory limit and therefore choose for articulatory reduction in terms of simplification of the articulatory gesture (cf. [Jannedy et al., 2010]). [Crystal and House, 1988] observed that this reduction was the case in more than half of the plosives they investigated which was in line with [Martinez et al., 1997]. Moreover, [Crystal and House, 1988] analyzed whether the position of a plosive within a word or syllable and its type of voicing influenced duration decrease. The authors discovered that the difference in plosive duration between several conditions was only marginal (5-10 ms). Here, voiceless plosives were the least affected. Additionally, [Martinez et al., 1997] found that affricates like [tʃ] were reduced to a pure fricative in fast speech. This was also discovered by [van Son and Pols, 1995], [van Son and Pols, 1996], [van Son and Pols, 1999]. In their investigations of the four ”spectral moments of fricatives” in normal and in fast speech they observed that a similar kind of weakening happened to fricatives in fast speech, too: The ”center of gravity” (henceforth ”CoG”) of the noise spectra investigated showed less intensity in fast speech. Provided that the CoG is defined as average frequency weighted by acoustic energy, it has a close connection to the vocal effort. The less intensity the CoG shows, the less vocal effort is invested, resulting in more reduction. [van Son and Pols, 1996] found that for all fricatives in spontaneous speech, the CoG was lower on average than the CoG for fricatives from read speech. This led to the conclusion that read speech was articulated with more vocal effort than spontaneous speech. These findings are in line with [Maniwa et al., 2009] who investigated the acoustic properties of clearly produced American English fricatives by analyzing their spectral moments as well. The researchers tested speech produced in an automatic speech recognition (ASR) scenario and elicited increased vocal efforts which were produced to emphasize putative misunderstood parts. They noted wide talker differences in types and in magnitude of the modification of speech and stated that there were consistent overall style effects observable for emphasized speech. Another phenomenon occurring in fast speech is the syllabification of consonants. Due to reduction and finally elision of vocalic segments, consonants may become the syllable nucleus. This is accompanied by a duration prolongation of the respective syllabified consonant [Kohler, 1990], [Roach et al., 1992]. [Roach et al., 1992] found that syllabic consonants [l], [n], [m] and [ŋ] were significantly longer on average than non-syllabic consonants across different speaking rates. They concluded that syllabic consonants are also important for the structure of an utterance and the perception of rhythm.

The proportion of durations as well as changes in degree of overlap in consonant clusters in fast speech were investigated by [Byrd and Tan, 1996]. The question under investigation was whether all consonants were reduced by the same amount when speaking rate was increased, or whether different consonants were affected differently as a function of consonantal class or position in the respective syllable. The authors noted that the rate of speech had a significant influence on the articulation of every single consonant combination they analyzed (cf. [Klatt, 1979]). The results were very speaker dependent, though, and sometimes even oppositional. In general, coarticulation was increased when speaking rate was increased, as was expected. Additionally, speakers with a higher habitual speaking rate also showed more articulatory overlap than speakers with a relatively slow habitual speaking rate. [Byrd and Tan, 1996] concluded that an almost linear proportion between speaking rate and degree of articulatory overlap existed for all speakers which also persisted, to a lesser degree, across individual realizations of utterances with different speaking rates by a single speaker.

Dynamic Features

Experiments on the mutual influences of vocalic and consonantal segments under different conditions are manifold, as well as examinations on fast speech effects. Many of those investigations have been presented in the previous section. However, listeners semm to be able to recognize a specific vowel even before target frequencies are reached [Lehiste, 1972], [Martinez et al., 1997] which is related to the acoustic information included not derivable from the static target formant frequencies but from other, dynamic features in vowel production. Therefore, this chapter examines dynamic features and their variation in fast speech. Transitions between vowels and consonants and vice versa will be discussed. Especially the changes of the Voice Onset Time (henceforth "VOT") in fast speech will be considered as it is important for the correct perception and identification of consonantal categories. In an early analysis, [Gay, 1978] found that the duration of transitions between vowels and consonants stayed relatively stable across different speaking rates. For slowly produced speech, the average transition duration was 40 ms to 50 ms, for fast speech it was 35 ms to 45 ms. This means that from slow to fast speech, transitions were shortened in duration by only 10% to 20% whereas the mid part of a vowel, where the formant frequencies were relatively stable, sometimes was completely elided. [Gay, 1978] pointed out that the transition durations he observed were shorter and less variable than the ones observed by [Peterson and Lehiste, 1960], [Lehiste and Peterson, 1961], and [Öhman, 1965], [Öhman, 1967]. The results of [Weismer and Berry, 2003], however, differ from those formulated by [Gay, 1978]. They discovered that the offset of the second formant of each vowel was largely influenced by changes in speaking rate, as well as by the duration of the following vowel. This observation was ascribed to coarticulatory effects which the authors claimed to only appear at a certain fast speaking rate. The importance of the preservation of the characteristic transitions of vowel formant frequencies to and from surrounding consonants, especially in fast speech, to ensure correct perception and identification of syllables or words by potential listeners was highlighted by [Martinez et al., 1997] as follows:

> ”It has been also observed that fast speech is basically a sequence of transitions from one sound to the next sound, so that the percentage of stable regions in the spectrogram is low in comparison with the same percentage at the average speech rate.” [Martinez et al., 1997].

The Voice Onset Time is another important acoustic cue for correct phoneme identification [Port and Dalby, 1982]. Before discussing the results of their own studies, [Kessinger and Blumstein, 1997] summarized the results of other investigations of the VOT as follows: Especially for American English, several evaluations had shown that changes in speaking rate had an influence on the VOT of certain consonantal categories in general, and that the outcome also had an impact on the perceptual category of sounds, especially in plosives ([Summerfield, 1981], [Miller and Baer, 1983], [Miller et al., 1986], [Miller and Volaitis, 1989], [Volaitis and Miller, 1992]; after [Kessinger and Blumstein, 1997]; cf. also [Miller et al., 1997], [Engstrand, 1988]). After [Kessinger and Blumstein, 1997] investigated the VOT of labial and alveolar plosives in detail, they found that the previously mentioned more general results were applicable to those specific consonantal categories as well. Additionally, they observed that the effect of shortening the VOT in fast speech was larger for voiceless plosives, as already noted by [Summerfield, 1981], [Port and Dalby, 1982], [Miller et al., 1986], and that the differences in VOT at the phonetic boundary between both categories became smaller when speaking rate was accelerated. The latter phenomenon led to an increased overlap between perceptual categories, similar to the one to observe for stressed-unstressed vowel pairs (cf. section 2.2.3), sometimes even neutralizing the VOT's function as a perceptual cue completely. An important aspect [Kessinger and Blumstein, 1997] pointed out in their conclusion was that the effects of changes in speaking rate were largely language-dependent and needed to be investigated for each language and phoneme in question separately.

2.2.3 Prosodic Organization

Speaking rate is an important characteristic of speech and as such closely linked to its *prosody* [Pfitzinger, 2001]. In general, the term ”prosody” refers to suprasegmental characteristics such as intonation, reflected by pitch and fundamental frequency; intensity, reflected by accentuation and stress; and quantity, reflected by duration. But also rhythm, phrasing and pausing are topics of interest in prosodic investigations [Bußmann, 1990]. [Lehiste, 1994] pointed out that supra-segmental characteristics could superimpose segmental characteristics but might not be limited to them. Consequently, supra-segmental characteristics reflect qualities of speech units larger than a phoneme, for example syllables, rhythmic feet, words, and phrases [Neppert and Petursson, 1986]. Next to supra-segmental features, prosody also includes extra- and paralinguistic phenomena such as speaker characteristics and the emotional or communicative situation [Möbius, 1995]. [Janse et al., 2003] pointed out that the role of prosodic factors becomes more important under difficult listening conditions because prosodic information is preserved better than segmental information ([Wingfield et al., 1984] after [Janse et al., 2003], cf. also [Sonntag, 1999]). Thus, although no consistent definition of prosody and its relation to supra-segmental characteristics is given in the literature [Kent and Read, 1992], the subsequent exemplifications rely on the above mentioned core areas of prosody assuming a generally accepted model of a phonological hierarchy of prosodic speech units.

Syllables are affected by an increase of speaking rate in terms of acoustic reduction and structural simplification. Anyway, not only syllable structure is simplified by leaving out certain phones ([Kuwabara, 1997], [Crystal and House, 1990]), and not only the acoustic characteristics of a syllable deteriorate in fast speech, but also the overall number of syllables in an utterance is reduced [Crystal and House, 1990], [Engstrand and Krull, 2001]. [Crystal and House, 1990] hypothesized that the articulation rate of a certain utterance was predictable from their syllable complexity. However, when analyzing their data they detected that the proportion of the average syllable duration between slow and fast speech was within the range of 2:1 to 3:1. The authors concluded that this proportion was speaker-specific and therefore, in contrast to their expectations, the articulation rate was not predictable from the syllable complexity of the produced utterance. [Kohler et al., 1981] came to a similar conclusion: Their hypothesis that syllables of different articulatory complexity would change their duration uniformly when speaking rate was accelerated was proven wrong by their own analysis.

[Gay, 1968] was one of the first researchers to evaluate changes in constituent duration of syllables when speaking rate was increased. He stated that all constituents of a syllable were shortened equally. Again [Kohler et al., 1981] revealed a similar result: The researchers noted that the duration of single segments was correlated with the overall syllable duration. However, changes in duration were neither linear nor did they show a relational regularity compared to changes in overall syllable duration. [Nooteboom, 1972], in contrast, found that the syllable-internal proportion of 1/3 consonantal and 2/3 vocalic parts stayed almost stable across different speaking rates. The *elasticity hypothesis* of [Campbell and Isard, 1991] stated that the relative durations of the syllable constituents were adjusted to the temporal frame of the syllable by scaling the intrinsic durations according to the temporal demands. Different factors were said to have an influence on this scaling, among them the number of phones in the syllable, the position of the syllable in the phrase, the stress assigned to the syllable, and the content of its parent word. However, in line with findings about vowel shortenings discussed above other investigations revealed that syllables resisted shortening beyond a certain point. [Fourakis, 1991], for example, stated that syllables as well as single vowels showed a tendency towards *incompressibility* (cf. [Klatt, 1979]). Also [Windmann et al., 2013] who developed an optimization-based model of speech timing based their investigations on this approach and suggested that incompressibility effects could be interpreted as a ”consequence of tradeoffs between competing requirements of production efficiency and communicative efficacy.” [Windmann et al., 2013].

The *principle of isochrony* presumes that constant time intervals between certain linguistic units exist. The nature of these linguistic units is different for different types of languages; particularly [Pike, 1945]'s early proposal to distinguish between stress-timed, syllable-timed, and mora-timed languages was adapted by other researchers (cf. [Delattre, 1966], [Abercrombie, 1967]). In stress-timed languages like German, isochrony is considered to apply to stressed syllables combined with a certain number of unstressed syllables, building a ”metrical foot” of the respective language. [Hoequist, 1983] hypothesized that any compression effects regarding isochronous intervals in accelerated speech would only appear in stress-timed languages. This assumption presumed that syllable durations were adapted to the shorter time frame available for metrical feet in fast speech. An elaborate investigation of isochrony in German was performed by [Kohler et al., 1981]. In addition to syllable complexity and the relation between segment and syllable duration, the authors evaluated the dependency of the realized overall temporal compression on the number of syllables contained in a metrical foot. Results showed that feet containing a different number of syllables with different complexity did not vary to the same degree when speaking rate changed. Foot length was rather dependent on the number of syllables and their complexity. However, no regular proportion was detected. There was only a slight tendency towards duration decrease of single syllables within a foot when syllable number was increased but this did not lead to isochronous intervals. These findings were supported by a follow-up study by [Kohler, 1983]. In this evaluation, speakers tended to choose a compromise between isochronous intervals and the prolongation of the foot duration in accordance with the number of contained syllables to be able to produce the requested speaking rate. In contrast, [Crystal and House, 1990]'s results revealed a constant, almost linear proportion between the duration of a stress group and the number of unstressed syllables being part of it. This observation was confirmed by [Brøndsted and Printz Madsen, 1997] with regard to the number of phonemes contained in a stress group. Additionally, [Crystal and House, 1990] found that long utterances were produced at faster speaking rates than short utterances. Similar results were presented by [Dankovičová, 1999] who observed that a word's articulation rate was significantly affected by its size in syllables: The rate was accelerated when the number of syllables contained in a word was increased whereas the overall articulation rate decreased throughout the intonation phrase. In addition, [van Santen, 1994] noted that speakers showed the tendency to speak faster in longer utterances. Moreover, unstressed syllables are generally shorter than stressed syllables in stress-timed languages [Crystal and House, 1988], [Gopal, 1990] which adds up to the observation that in fast speech unstressed syllables are reduced and shortened more than stressed syllables in general. [Dellwo and Wagner, 2003] conducted several experiments on the influence of speaking rate on vocalic and intervocalic measures for different languages. They assumed that the vocalic and intervocalic measures %V and ΔC distinguished between linguistic rhythm classes reflected in varying proportions of variance for %V and ΔC in relation to speech rate within and across languages. The authors analyzed the percentage of vocalic intervals %V and the standard deviation of consonantal intervals ΔC in the respective speech signals of three different languages produced at five different speech rates. While ΔC showed a significant influence of speech rate, %V remained almost constant across different speaking rates. Based on their findings, the authors proposed a new model of tempo control in speech based on the "intended speech rate" (ISR) instead of the objectively measurable "laboratory speech rate" (LSR). However, constancy of time intervals as assumed by the isochrony principle was seldom verified, and if then only for short stretches of speech [Lehiste, 1977], [Dauer, 1983]. Therefore, this approach is still discussed controversially in the literature.

Other prosodic features like *intonation* and *pitch* are influenced by an acceleration of speaking rate as well. Here, the term "intonation" refers to the rise and fall of the voice pitch over entire phrases and sentences, and is acoustically reflected in the fundamental frequency f0. "Pitch", however, refers to the perception of the relative highness or lowness of a tone. It is the perceptual correlate of the fundamental frequency F0. In an early investigation, [Cooper et al., 1983] conducted an acoustic analysis of the fundamental frequency characteristics of habitually fast versus habitually slow speakers. They found that the fundamental frequency showed somewhat higher F0 peaks for habitually fast speakers as well as for fast rates of speech. The researchers ascribed this finding to the increase in muscular tension when speaking rate was accelerated. In general, they found steeper F0 slopes in fast speech which they attributed to a compensatory effect, as the same movements needed to be carried out in shorter time when speaking rate was accelerated. Also [Fougeron and Jun, 1998] found speaker specific effects on the phonetic realization of the F0 contour in addition to more general influences of the acceleration of speaking rate. In their investigation, the increase in speech rate caused a reduction in pitch range as well as pitch excursion less pronounced in different parts of a recorded text for different speakers. The tonal contour of utterances was simplified, the overall intonation contour thus became flatter (cf. [Beaugendre, 1995], [Monaghan, 2001]). Due to its monotony, "this speaking style can give the listener the impression of tediousness" [Fougeron and Jun, 1998]. Looking at the strategies applied by various speakers to accelerate speaking rate it became obvious that different speakers tended to lower the F0 maxima and/or raise F0 minima to a different extent. A similar observation was made by [Trouvain and Grice, 1999] who noted that changes in F0 were very idiosyncratic and did not show systematic patterns across speakers.

The evaluation conducted by [Caspers and van Heuven, 1991] focused on accent lending pitch movements in fast speech in Dutch. The researcher found that speakers did not economize on accent lending pitch movements when accelerating speech rate, but noted that 40% of boundary marking pitch movements were simplified or even left out. From this observation, they derived the conclusion that there must be a difference between obligatory and optional boundary marking pitch movements. In a follow-up study, [Caspers and van Heuven, 1992] investigated the excursion size, the duration and the steepness of pitch movements elicited under time pressure. She found that the pitch rises became shorter in terms of duration but not in terms of frequency range, and that pitch movements were steepened (cf. also [Cooper et al., 1983]). The time compression was highest for multiple pitch movements within the same time span. [Caspers and van Heuven, 1992] summarized her findings in stating that "[c]ertain movements have fixed anchor points relative to the segmental structure, other movements are free to range, but only within a limited domain." Especially the correct timing of the beginning of the pitch rise showed to be important. This was also found by [Prieto and Torreira, 2007] for the alignment of LH* pre-nuclear peaks in Castilian Spanish: Peaks were located later in the syllable if speaking rate was increased for both open and closed syllables. However, the authors noted that in general gestures at syllable onsets seemed to be more tightly coordinated than gestures at syllable ends which was in accordance with the segmental anchoring hypothesis for tonal landmarks as well as previous findings discussed above (cf. [Greisbach, 1991], [Caspers and van Heuven, 1992]). Nevertheless, the authors pointed out that their findings may have been very language dependent as other research showed that the alignment of the H tone was not affected by speech rate to the same extent as in their data. Also [Trouvain and Grice, 1999] revealed a reduction of the overall pitch range as well as the amplitude of rising and falling pitch movements in addition to a simplification of tonal structure in fast speech. Nonetheless, their findings were again not applicable to all speakers: Some realized noticeably less pitch accents when speaking faster while others showed only slight differences. Just one speaker used more monotonal accents in fast speech compared to more bitonal accents in normal speech, and different strategies were applied again when speaking rate was slowed down. [Wu and Sun, 2000] investigated fast alternating high and low pitch sequences elicited from native speakers of Mandarin and English who were not singers. They measured the time needed to complete the middle 75% pitch shift ("response time") and the time needed to complete the entire pitch shift ("excursion time") and detected that the latter was almost twice as long as the first. The researchers concluded that the maximum change of pitch speed was not nearly as fast as previous data implied because of severe physiological limitations. With regard to speaker origin, they discovered that the excursion time was longer for English speakers, but the speed of pitch change was faster than for Mandarin speakers because of the larger excursion size.

An increase in speech rate does not only affect phrasing and pausing [Caspers and van Heuven, 1991], [Fougeron and Jun, 1998], [Monaghan, 2001], but can also have an influence on the prosodic organization of a complete text. [Fougeron and Jun, 1998] observed that this phenomenon varied according to speaker and position of the speech material in the text. Especially minor phrase boundaries are affected, followed by to initial accents of which only 25% remained in fast speech [Beaugendre, 1995]. [Caspers and van Heuven, 1991] found that 40% of boundary marking pitch movements were left out in their data. They concluded that there must be a difference between obligatory and optional intonation phrases. In accordance with [Trouvain and Grice, 1999], the authors claimed that intonation phrases can be restructured and boundary marking pitch movements can be simplified in fast speech. However, in contrast [Trouvain and Grice, 1999] pointed out that despite a notable reduction of number and strength of phrase boundaries the reduction of the number of prosodic breaks in terms of their ToBI analysis was only moderate in fast speech. Nevertheless, breaks were substantially demoted when speeding up. [Keller and Zellner, 1996] noted that in their experiments speakers tended to sacrifice some "niceties" of phrase-internal timing modulation in considerably accelerated speech, only keeping phrase-final durational markers. The number and duration of pauses in normal versus fast as well as fast and clear speech were evaluated for the selected speaker; results are discussed in chapter 6.2.1.

2.3 Speaking Styles and Speaking Strategies

Despite the continuous flow of speech accompanied by coarticulation, a sufficient contrast between neighboring segments is both necessary and achievable in successful human communication. According to Lindblom's theory of hyper- and hypoarticulation ("H&H theory") [Lindblom, 1990], a contrast is sufficient if it allows the listener to discriminate the signal to the extent necessary to identify the intended item in his mental lexicon. On the other hand, a speaker aims at the production of earmarked and future-oriented speech. This causes a dilemma because a speaker tries to communicate with as little effort as possible which results in "hypospeech", a somewhat more slurry pronunciation style. But as the very same speaker also wants to reach a communicative goal s/he needs to maintain the phonetic contrast necessary for comprehension. Thus, in situations where comprehension might be more difficult (for example in a loud environment) or absolutely essential (for example when giving instructions) speakers tend to use "hyperspeech", a very exact and pronounced speaking style. Lindblom describes this phenomenon as follows: "[S]peakers are expected to vary their output along a continuum of hyper- and hypospeech." [Lindblom, 1990]. In a follow-up study, [Lindblom, 1996] noted that in fast (hypo-)speech large displacements took place which he called "undershoot effects" (cf. section 2.2.1), observable especially at the midpoint of short vowels. In contrast, when vowel duration was prolonged the target frequencies of the respective formants were approached more and more.

[Engstrand, 1988], however, came to a different conclusion since the data he evaluated failed to produce evidence for an undershoot mechanism. None of the spectral characteristics of the analyzed vowels were significantly influenced by changes of the speaking rate. He attributed this observation to an active temporal restructuring of articulation and concluded that speech tempo as well as speaking style were controlled independently (cf. [Zwicky, 1972]). Furthermore, speakers seemed to have the option of either decreasing articulatory movement amplitude or avoiding undershoot by means of an increase in movement velocity at higher speaking rates (cf. [Kuehn and Moll, 1976]). Moreover, [Engstrand, 1988] argues that [Lindblom, 1983] had shown that the excursion of articulatory movements may depend on both duration and vocal effort. Therefore, he holds that it is conceivable to interpret the absence of undershoot in his data as "the subjects' use of greater articulatory precision under fast and stressed speaking condition as compared with the slow or unstressed conditions." [Engstrand, 1988]. These findings were supported by [Pols and van Son, 1993] who investigated static and dynamic formant characteristics of vowel segments and found no indication for duration-dependent undershoot. Speakers rather adapted their speaking style to the requested speaking rate to reach the same midpoint formant frequencies as in normal speech. [van Bergem, 1993] resumes that reduction in fast speech "is most likely dependent on the particular speech style a certain speaker uses and not on physiological constraints of his articulatory organs." [van Bergem, 1993]. In line with those findings, [Bradlow et al., 1995] and [Bradlow, 2002] found that a substantial portion of variability in normal speech intelligibility was traceable to specific acoustic-phonetic characteristics of the talker. Especially the expansion of the vowel space was correlated with the intelligibility of the speech produced.

In addition, it can be observed that the degree to which speakers vary their speaking rate is very different [Barry, 1998], [Mixdorff et al., 2005], [Mok, 2007]. [Trouvain and Grice, 1999], for example, noted that speakers differed in the extent to which articulation rates where changed across subjective speaking rates "slow", "normal", and "fast". The observation that differences between speaking rates were bigger for measured speaking rate than for articulation rate was mainly attributable to differences occurring in pausing. [Gay, 1978] found out that it was the most difficult condition for speakers to have a free choice in speaking rate. Neither were subjects able to produce two different fast speech rates, nor was slow speech produced consistently. [Martinez et al., 1997] detected that self-defined speech rate groups showed a significant overlap in absolute rate of speech. The researchers concluded that the selection of a certain speaking rate was highly subjective. However, [Dellwo and Wagner, 2003] came to a different conclusion. They deduced from their evaluation of speech produced at five different speaking rates that speakers may have a notion about normal, slow, and fast speech rate in their language. The number of syllables per second which a speaker of a certain language was able to produce showed to be language dependent. Therefore, the authors proposed to use the "intended speech rate as basis of speech rate control" [Dellwo and Wagner, 2003] instead of the "laboratory measurable speech rate".

[Engstrand, 1988] also investigated the relation between speech tempo and speaking style. He introduced the terms "careful" as opposed to "casual" speech when looking at the speaking style, and stated that speaking style was independent from speech rate. In earlier investigations, researchers were not in complete agreement with this. [Dressler, 1972], for example, noted that fast speech was the most careless speaking style and therefore equal to casual speech. In a later study, also [Engstrand and Krull, 2001] defined fast speech as being casual. [Fosler-Lussier and Morgan, 1998] stated that spontaneous speech was clearly different from fast read speech. They found more word pronunciations deriving from the canonical form in fast speech compared to normal speaking style. The phoneme deletion rate increased from 9.3% to 13.6% in very fast speech, and the entropy of the distribution of pronunciations grew as well, implying more variability in fast speech. [Mefferd and Green, 2010] found that the intelligibility of speech was higher with less variability. To reach higher intelligibility, speakers were expected to apply different articulatory strategies when producing (fast) speech. In their experiments, the researchers revealed talker specific articulatory responses to speaking rate and loudness manipulations. They concluded that only speaking rate reduction may be an effective articulatory strategy to enhance speech intelligibility. [Hasegawa, 1979], however, stated that fast speech does not per se exclude clear speech. Moreover, he declares that fast speech does not necessarily equal casual speech, but that casual speech is mostly deployed dependent from the situational context.

This was not only noted by [Lindblom, 1990] in his theory of hyper- and hypo-articulation as well, but also by [Maniwa et al., 2009] and [Kuehn and Moll, 1976] who observed that speakers were able to adopt a speaking style that allowed them to be understood more easily in difficult communication situations by applying different speaking strategies. In parallel to [Barry, 1998] and [Mok, 2007], [Liu and Zeng, 2006] as well as [Maniwa et al., 2009] found wide talker differences in types and magnitude of modification when clear speech was produced as opposed to conversational speech. The authors determined that clear speech had an intelligibility advantage of up to 38% compared to conversational speech for hearing impaired listeners as well as in noisy environments (cf. [Krause and Braida, 2002]). They observed that clear speech involved less frequent vowel reduction, more and longer pauses, a reduced speaking rate, an increased mean and range of fundamental frequency, and an expanded vowel space. Additionally, voicing contrasts and place of articulation contrasts seemed to be enhanced in clear speech. Consonants, however, especially non-sibilants, were a large source of listening errors. To ensure high intelligibility of speech in general it showed to be important to verify that especially consonants were highly intelligible. The greatest benefits in intelligibility were observed for speakers who produced fricatives with a much larger energy in higher frequencies.

Referring to [Lindblom, 1963] and [Lindblom, 1990], [Moon and Lindblom, 1994] investigated formant patterns of vowels produced in clear speech elicited by applying the feedback method in noise where speakers increased their vocal effort. He found that ”clear samples were not merely louder, but involved a systematic, undershoot-compensating reorganization of the acoustic patterns.” [Moon and Lindblom, 1994]. Formant patterns’ displacements were mostly dependent on vowel duration and context. That way, he developed a revised, bio-mechanically motivated version of the undershoot model proposed by [Lindblom, 1990]. Also [Amano-Kusumoto and Hosom, 2010] conducted an evaluation of formant contours for clear versus conversational speech. They calculated coarticulation coefficients for each speaking style and revealed that those were different for different speaking styles. Additionally, they noted that formant targets were independent from the respective speaking style. Only slopes at the vowel onset differed in steepness. The movements of the articulators were faster when clear speech was produced in contrast to conversational speech. According to [Bradlow et al., 1995] and [Maniwa et al., 2009], a more clear articulation would be accompanied by an expanded vowel space and a higher vowel space dispersion compared to a more conversational speaking style, too. [Krause and Braida, 2002] found that clear speech was best elicited by repeated pronunciation of the same utterance by experienced speakers. They concluded that the advantages of clear speech could be extended to slightly faster speaking rates but maybe not to quick speech because the ”clear and quick” samples they evaluated were significantly slower than ”conversational and quick” utterances. The authors also assumed that clear speech may have had some inherent properties that made it more intelligible. Thus, in line with previous findings, also those researchers stated that ”clear speech is a speaking style that speakers adopt in order to be understood more easily in difficult communication situations.” [Krause and Braida, 2002]. (Cf. also [Grice, 1989] and [Pols, 1999]).

The ability of a speaker to speak fast and clear at the same time is an important aspect of intelligibility and overall speaking style. In this context, a certain degree of habitually clear speaking style proposes additional advantages [Greisbach, 1992], [Liu and Zeng, 2006]. In line with the findings of [Krause and Braida, 2002], [Jannedy et al., 2010] hypothesized that the highly repetitive training of specific sequences removes linguistic planning as the speaker learns to repeat articulatory templates. Therefore, this approach was applied when corpus recordings for the current project were conducted, as explained in chapter 7.1. Moreover, [Trouvain et al., 2008] found that female as well as male voices with enlarged pitch range and faster articulation rate than the default setting of the diphone synthesis investigated in their evaluation got better overall impression scores. Additionally, faster speakers appeared more competent and convincing, more confident, more intelligent, and more objective ([Smith et al., 1975] after [Trouvain et al., 2008]). The implications of these findings for defining a suitable speaker as well as the specific characteristics of this speaker who will be selected to record a fast speech unit selection inventory will be presented later on in chapter 6.

2.4 Summary and Conclusions

The characteristics of natural fast speech have to be taken into consideration when investigating the modeling of fast speech in unit selection speech synthesis. Therefore, in the current chapter different aspects of natural fast speech production were examined. At first, the definition of *speaking rate* was outlined in section 2.1. Different approaches to measure speaking rate were explained afterwards. In particular different units of measurement were discussed. The approach to determine speaking rate for the speech material to be produced and investigated in the work presented here was defined. Measurements of speaking rate as presented in chapters 6.2.1, 7.1, and 8.2 will be done in syllables per second for single utterances. This method was chosen since natural speech inherently includes a certain variability of speaking rate. Corpus recordings as well as target utterances generated with the specific speech synthesis system applied had to fulfill a predefined range of speaking rate per utterance, but a detailed, fine-grained definition like the local speaking rate was not considered necessary.

The manifestation of changes in speaking rate as well as their effects on different linguistic units were described subsequently in section 2.2. It was pointed out that fast speech differed significantly from speech produced at a normal speech tempo both in articulatory and acoustic characteristics. At the beginning, the terms *coarticulation* and *reduction* were discussed from a general perspective. Their impact on articulation was examined in section 2.2.2. Since articulation in fast speech has to take place in a smaller time frame, linguistic units are usually produced with more gestural overlap and acoustic interference [Davidson, 2006]. [Jannedy et al., 2010], however, observed that articulatory movement amplitudes remained still very large in fast speech for their subject, a highly trained speaker. The authors concluded that articulatory reorganization as well as speech errors were avoided by means of training of repeated patterns (cf. also [Greisbach, 1992], [Liu and Zeng, 2006]). This hypothesis was the basic principle of the procedure applied during corpus recordings outlined in chapter 7.1: To approach the fastest speaking rate possible, the speaker generally followed the strategy of repeating accelerated renditions of a sentence several times in a row.

Acoustic alterations of characteristics of single speech segments as well as of transitions between them were detailed afterwards in sections 2.2.2 and 2.2.2. Acoustic alterations of single segments mostly apply to overall duration and most characteristic acoustic features, like formant frequencies for vowels or spectral moments for consonants. Since the fast speech produced by the selected speaker will be analyzed acoustically (cf. chapter 6.2), certain aspects of shortening and reducing vocalic and consonantal segments were explained in more detail in the current chapter. However, these phenomena are generally not desirable when creating a fast speech corpus to be applied in a unit selection synthesis system. Shortening in duration, for example, might occur to different extent in fast speech and lead to complete elision of segments. Since next to vowels also different classes of consonants were found to be shortened in varying ways and to a different degree when speaking rate increased (cf. section 2.2.2), the duration differences between the different speaking styles elicited from the selected speaker will be examined separately for vowels and single consonants as well. Consonantal segments will be grouped by manner of articulation for further analysis (cf. section 6.2.1). To verify that the selected speaker is able to preserve certain acoustic contrasts important for correct perception in fast and clear speech (cf. [Kessinger and Blumstein, 1997]), acoustic reduction of vocalic segments in terms of changes in formant frequencies and vowel space (cf. [Fourakis, 1991]) will also be evaluated (section 2.2.2). An analysis of the observable acoustic reduction in terms of overall acoustic differences by means of spectral similarity between the different realizations of fast speech will complete the comparative acoustic evaluation of the selected speaker's normal and fast speech (cf. section 6.2.1). Implications of changes in speaking rate for larger linguistic units such as syllables, words, and phrases were discussed afterwards in section 2.2.3. For those linguistic units, only the number and duration of pauses will be investigated later on (cf. also section 6.2.1). A more detailed evaluation of other prosodic features of fast speech produced by the selected speaker would go beyond the scope of the current work, but might be subject to further investigations in the future.

The last section of this chapter (section 2.3) dealt with speaking strategies applied to produce different speaking styles. As outlined before, speakers mostly follow certain strategies when speaking fast: Vowels and consonants are acoustically reduced and shortened in duration, the fundamental frequency contour is flattened, and the duration of pauses is minimized. These phenomena may lead to a loss of distinctiveness in speech and consequently a loss of comprehension on listeners' side. However, speakers obey certain rules in order to keep the communication chain working: Important elements of speech are reduced less than unimportant ones. With additional effort, speakers are well able to speak both clear and fast (cf. for example [Lindblom, 1990]). This specific speaking style, namely the production of fast and clear speech, was seen as desirable for the current research since it was assumed to be suitable for use in a fast speech unit selection synthesis system. From this assumption, specific requirements for the selection of a suitable speaker will be defined in chapter

6. It is expected that modeling fast and clear speech in unit selection speech synthesis may increase the naturalness of synthetic speech without harming its intelligibility. Also with regard to the definition of the adequate unit size for fast speech synthesis (cf. chapter 8.1), this aspect plays an important role. The clearer fast speech is produced, the more it is possible to preserve important acoustic cues also in speech units to be used in a concatenative speech synthesis. Thus, the aim of the work presented here was to integrate the insights of [Lindblom, 1990] and a more flexible approach to inventory creation for unit selection synthesis system in order to achieve synthetic speech that is both maximally natural and maximally fast.

Chapter 3

Modeling Speaking Rate in Speech Synthesis

Research in speech synthesis can be based on different motivations. Mostly, either the speech production process or the acoustic characteristics of speech are of main interest. Another motivation is the aim for a most natural and optimal speech synthesis output. Depending on the goal of research, the synthesis technique applied plays a crucial role. Articulatory synthesis, for example, is most suitable to investigate speech production and articulation processes. If, however, the generation of natural sounding speech is the target of investigations concatenative unit selection synthesis techniques will be the preferred choice as they are based on the approach to put together units cut from natural speech to generate the desired utterance. Concatenative unit selection speech synthesis can either rely on uniform speech units like half- or diphones whose duration and other acoustic parameters are manipulated to meet the criteria defined by the utterance to be synthesized, or be a true unit selection synthesis in terms of selecting suitable units of different size from an underlying unit selection corpus. Different approaches of speech synthesis are outlined in section 3.1. The perceived naturalness of synthetic speech is notably enhanced by an adequate duration prediction [Brinckmann and Trouvain, 2003]. The duration of speech segments is affected by many different factors. When implementing fast speech as a unit selection corpus in speech synthesis, the possibility as well as necessity to adapt the duration prediction module are therefore of main interest. Thus, different approaches of duration prediction in speech synthesis are discussed in section 3.2.1 before methods and algorithms to accelerate artificially generated speech to even higher speaking rates than natural ones are examined in section 3.2.2.

3.1 Synthesis Techniques

The actual synthesis of speech is only the last step of a chain of processing steps implemented in a text-to-speech (TTS) system. The initial processing steps dealing with the conversion of the incoming text into phonetic symbols as well as adequate prosody generation are independent of the actual speech synthesis approach. Since none of these linguistic components are of prior interest for the current research, only different techniques to synthesize speech are explained in the following paragraphs. Classical speech synthesis approaches are divided into parametric approaches and data-driven approaches [Hess, 1994], [Sproat and Olive, 1995]. The main features of parametric synthesis like formant synthesis or articulatory synthesis will be opposed to data-driven approaches like synthesis from large corpora and multi-level non-uniform unit selection. Especially the latter is examined in more detail, as the research conducted for this work was based on the open source speech synthesis system developed at the University of Bonn (Bonn Open Speech Synthesis, BOSS, [Klabbers et al., 2001], [Breuer and Hess, 2010]).

The phenomenon of coarticulation (cf. section 2.2.1) is a good example to illustrate the differences between parametric and data-driven synthesis techniques. Coarticulation is a dynamic effect emerging during the articulation process when articulator movements consecutively executed to reach articulatory targets are smoothly concatenated. Such smooth transitions are crucial for the intelligibility of natural as well as synthetic speech (cf. chapter 4). During parametric synthesis, such transitions are modeled by applying suitable parameters. Those parameters usually are adjusted by a set of rules in formant synthesis, for example, or by applying a dynamic model like in articulatory synthesis. Therefore, the main task for parametric synthesis research is to define an appropriate set of parameters for specific phones and phone combinations. They are often defined during an elaborate "trial-and-error" procedure where initial values are derived from an extensive evaluation of acoustic parameters of existing speech data (cf. [Elsendoorn, 1985]). In contrast, data-driven approaches reuse predefined parts of recorded speech cut from the original speech signal. Here, coarticulation is covered by units inherently comprising the desired phone transitions. This is the case for diphones, for example, which are defined to start in the (more or less) stable middle of a phone and to end in the (more or less) stable middle of the following phone. Research on data-driven speech synthesis is mainly focused on finding the optimal speech units for concatenation as well as defining and optimizing the unit selection process, mainly in terms of cost functions to be applied. The former goal is also worked in the current investigations as outlined in chapter 8.1 whereas the latter is beyond the present research.

3.1.1 Parametric Synthesis

The foremost representatives of parametric speech synthesis techniques are formant synthesis and articulatory synthesis. More recently, also statistic parametric synthesis approaches based on HMM models have been developed. While formant synthesis is based on the acoustic analysis of natural speech data, articulatory synthesis uses parameters derived from the mathematical description of articulator movements. The following explanations about differences between those approaches widely follow [Möhler, 1998].

Formant synthesis relies on the source-filter model as underlying principle for the generation of artificial speech [Fant, 1960]. In this model, speech generation often is divided into two separate processes: A linear filter that models the vocal tract, and a specific source that stimulates it. The two parts - the source on the one hand and the filter on the other hand - are assumed to be independent from each other. In human speech production, voiced excitation happens through excitation of the vocal folds. This can be modeled by a pulse source with low-pass characteristics. The frequency response of the vocal tract is characterized by the occurrence of specific formant frequencies. They result from the amplification of the excitation wave deriving from the larynx in the vocal tract. Thus, formant frequencies appear as maxima of the frequency response of the vocal tract. To enhance the naturalness of the generated signal, parametric and intonational models are applied [Klatt and Klatt, 1990]. Voiceless sounds, however, are produced with open vocal folds in human speech production. Turbulent air flows induced by constrictions as well as complete closures of the vocal tract are responsible for the emergence of fricatives and plosives. To synthesize such kind of sounds, usually a (pseudo-)random generator is applied to produce white noise. During nasalized articulation, the nasal cavity is connected to the actual vocal tract by lowering the velum. This results in turn in an attenuation of the excitation wave manifesting itself as anti-formants or zero points in the frequency response characteristics of the vocal tract. Hence, the frequency response to be modeled contains maxima based on the formant frequencies as well as zero points derived from the anti-formants. In formant synthesis, extrema are generated by band-pass filters and zero points by filter attenuation bands. The speech synthesis system ”JAWS Eloquence” provided by [FreedomScientific, 2011] which was used as a reference system to compare with during the evaluation of the fast speech generated with the unit selection synthesis system BOSS (cf. chapter 8) is such a formant synthesis system.

Where formant synthesis is referring to acoustic measurements while taking into account some phonological constraints, articulatory synthesis is based on the knowledge of human speech production processes. According to [Birkholz, 2005], in contrast to formant synthesizers which specify the formant frequencies and bandwidths as well as the source parameters directly, articulatory synthesizers determine the characteristics of the vocal tract filter by means of a description of the vocal tract geometry, and place the potential sound sources within this geometry. Different kinds of articulatory synthesizers reproduce the vocal tract geometry either in one, two or three dimensions. In a one-dimensional model, the vocal tract is described by means of its area function. This function describes the change of the cross sectional area of the vocal tract between glottis and mouth opening over time. [Birkholz, 2005] explicates that ”[a]ssuming one-dimensional sound propagation in the vocal tract, the area function contains all information to specify the filter characteristics.” This is also the reason why two- and three-dimensional vocal tract models are converted into a one-dimensional area function for acoustic simulation. However, such multi-dimensional models allow for much better and more direct simulation of the position and form of the articulators involved. The models are shaped by means of a small set of articulatory parameters. Again, their variation over time influences the changes of the vocal tract area represented by the area function. Based on the sequence of area functions and their corresponding sound sources, an acoustic model is used to estimate the resulting waveform. [Birkholz, 2005] summarizes that according to previous explanations, an articulatory synthesis system always comprises at least three components: ”[A] geometric description of the vocal tract based on a set of articulatory parameters, a mechanism to control the parameters during an utterance, [and] a model for the acoustic simulation including the generation of the sound sources”. Deeper matters can also be found in [Birkholz, 2016].

The third approach to be categorized as parametric speech synthesis is the so called ”HMM-based speech synthesis”. Originally, the usage of Hidden-Markov-Models (HMMs) was common in Automatic Speech Recognition (ASR) but over time they found their way into speech synthesis as well. HMM-based synthesis is also called ”statistical parametric synthesis”. In such systems, the frequency spectrum reflecting the vocal tract shape, the fundamental frequency deriving from the voice source, and the duration influencing the prosody are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion. [Black et al., 2007] gave a detailed overview over techniques applied in statistical parametric speech synthesis. In their exemplifications the authors elaborated HMM-based speech synthesis in contrast to the more conventional approach of unit selection synthesis, and discussed their advantages and disadvantages. According to them, an HMM-based synthesis system consists of two main parts where the training part is similar to the one applied in speech recognition systems. However, the main difference is that both spectrum and excitation parameters are extracted from a speech database and modeled by context-dependent HMMs. Phonetic, linguistic, and prosodic contexts are thereby taken into account. Thus, HMM-based speech synthesis models spectrum, excitation, and duration in a unified framework. The second part of the system is the synthesis component. Corresponding to [Black et al., 2007], an HMM synthesis system can be roughly described as follows:

> [F]irst, a text is converted to a context-dependent label sequence and then the utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. Secondly, state durations of the HMM are determined based on the state duration probability density functions. Thirdly, the speech parameter generation algorithm generates the sequence of mel-

> cepstral coefficients and log F0 values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated mel-cepstral coefficients and F0 values with binary pulse or noise excitation. [Black et al., 2007].

A more elaborate description can also be found in [Zen et al., 2007].

To sum up, [Black et al., 2007] note that statistical parametric synthesis might be most simply described as ”generating the average of some set of similarly sounding speech segments” whereas unit selection speech synthesis sounds more natural on average. However, according to the authors the unit selection approach has some other major disadvantage: The quality of the synthesized speech declines severely if an utterance is not or only partially covered by the inventory. In contrast, statistical parametric synthesis allows for the generation of any average utterance, although the output often sounds less natural than the one of a unit selection synthesis system, and the reconstruction process from parameters is still not ideal. Generally speaking, the results of several comparative listening evaluations showed that HMM-based speech synthesis seemed to be preferred and better to understand, but best examples - in terms of most natural sounding utterances - were still originating from unit selection synthesis (cf. [Black and Taylor, 1997]). Moreover, the authors point out that they observed three main factors degrading the quality of speech synthesized when deploying statistical parametric synthesis. The first was the Vocoder which introduced some buzziness, secondly, the modeling accuracy may have been inadequate, and thirdly over-smoothing could have made the generated speech sound muffled. [Black et al., 2007] conclude their explanations with an accentuation of the advantages of the HMM-based generation synthesis approach which in particular are easy voice characteristic modification, applicability to different languages requiring only small adaptations, and variability. Additionally, techniques applied in ASR could easily be adopted for speech synthesis, and the footprint of the technology is relatively small. On top of that, [Zen et al., 2007] point out that speech synthesized by unit selection synthesis is limited to the style of the speech recorded in the database whereas HMMs are only trained from a database of natural speech. Therefore, an HMM synthesis system offers the ability to model different speaking styles without requiring the recording and preparation of very large natural speech databases. It will be discussed in chapter 9 whether statistical parametric synthesis could be an option to model fast speech in speech synthesis as well.

3.1.2 Data-driven Approaches

In contrast to parametric approaches where the speech signal is generated completely artificially, data-driven approaches rely on speech units derived from a corpus of natural language utterances. Units are cut from this speech corpus and concatenated during synthesis to form the required utterance. The quality of the synthesized speech largely depends on the quality and quantity of the underlying corpus as well as on the scope of the application. Most current unit selection speech synthesis systems were developed to generate texts from almost any domain. However, there are also systems available which were designed for a very small, closed domain. Those contain almost all domain-specific words or phrases in the corpus. Therefore, this approach is also called ”canned speech synthesis”. Such systems are applied in public transport, for example, to announce stations, or in fixed dialogue systems. Any extension requires additional speech recordings to be implemented. A system design between canned and open domain systems is most suitable for a restricted domain [Möbius, 2003].

In his early research about synthesized speech, [Harris, 1953a] (after [Breuer, 2009]) realized that the generated speech was unintelligible when concatenating distinct phonemes. Also his investigations into ”building blocks of speech” revealed unsatisfying artificial speech quality, although he already had realized that several allophones per phoneme were desirable [Harris, 1953b] (after [Breuer, 2009]). The bad outcome was due to the fact that transitions from one sound to the other are very important for the intelligibility of speech (cf. section 2.2.2), but were not included in the single distinct phone units. Transitions between phones should have been kept to enhance intelligibility. Relying on this finding, [Peterson et al., 1958] developed the concept of ”diphones”. Diphones were defined as ranging from the central point of an almost static spectral state of one phone to the center of an almost static spectral state of the subsequent phone. Hence, the transition from one sound to the other was incorporated in the diphone. To date, diphones are one of the basic units applied in concatenative speech synthesis as they allow for a quite natural and intelligible speech generation based on the transitional information they contain [Moulines and Charpentier, 1990], [Pols, 1992], [Sproat and Olive, 1995]. Also, the size of the required unit inventory is rather small [Portele, 1996]. For German, for example, [Portele, 1996] found that approximately 2000 diphones would suffice to include all possible diphone combinations in German. Still, since diphones only take into account the direct context of a phone, they do not cover for long-range coarticulatory phenomena occurring in natural speech. Hence, the set of speech units used for synthesis of more complex languages where consonant clusters influencing each other appeared more often (like for example in German) was extended: [Fujimura and Lovins, 1978] as well as [Ruske and Schotola, 1978] (after [Breuer, 2009]) invented the concept of ”demi-syllables” which spanned either the onset of a syllable and part of the nucleus until the almost static spectral center, or the second half of a syllable starting in the static spectral center of the nucleus and ending right after the end of the coda. This way, nearly all coarticulatory effects occurring within syllables were covered. The increasing size of the required inventory, however, was problematic. According to [Dettweiler, 1984], for German 5400 demi-syllables were needed. However, despite general improvements also demi-syllables could not cover for more extensive coarticulatory phenomena spanning more than one syllable.

The increased size of the inventory was the reason why [Portele, 1996] eventually argued for a mixed inventory, consisting of a set of demi-syllables covering not only for phonological syllables but also for a more phonetic-acoustically motivated syllable structure, plus some additional units of different size to cover for missing phones (for example coda obstruents) or rather complex consonant combinations. The units he defined were taking into account the characteristics of the preceding syllable if coarticulation was expected as well as the possibility of adding final obstruents separately. Thus, the inventory [Portele, 1996] proposed for German contained approximately 2200 units instead of 5400 demi-syllables.

Another approach considering long-range coarticulation was described by [Olive, 1990]. The researcher suggested to include triphones and short words in the inventory. However, neither this method nor any of the other ones explained before were designed to factor in even more extensive, long-range coarticulation across syllables or larger speech units like words. Moreover, the large number of unit concatenations necessary in diphone synthesis adversely affected the quality, especially the naturalness of the synthesized speech [Möbius, 2000]. Therefore, the core idea of corpus-based unit selection was developed, leading away from pure concatenative approaches to synthesis from large corpora: To select the longest possible string of phonetic segments at runtime to minimize the number of concatenations and reduce the need for signal processing. According to [Möbius, 2000], the complexity and combinatorics of language and speech were posing the main challenge here. The relative weighting of acoustic distance measures as well as the development of appropriate criteria to create a unit selection inventory with optimal coverage of the target domain were additional issues to solve. For those who would like to get to know more about the evolution of non-uniform unit selection speech synthesis it is recommended to read the full article [Möbius, 2000].

Speech synthesis from large corpora

Despite the aforementioned challenges, [Sagisaka, 1988] (after [Breuer, 2009]) was the first one to propose non-uniform unit concatenation in 1988. When developing "CHATR", the speech synthesis system of the ATR in Kyoto, Japan, [Black and Taylor, 1994] and [?] (after [Breuer, 2009]) finally overrode the until then common approach of uniform or at least well defined unit sizes and deployed a complete speech corpus as unit selection inventory. A very efficient search algorithm looked for the most suitable units contained in the corpus and concatenated them during runtime. Here, the size of the unit was not predefined. Unit sizes ranged from very small demi-phones to the longest string possible. Most of the time, several neighboring phones were put together to generate the desired utterance. Also the exact point of concatenation was not predefined but determined during run time. According to [Campbell, 1996], this non-uniform unit selection approach was to be seen as a reorganization of the speech units comprised in the corpus:

> It is customary to consider source units as an integral part of the synthesiser, but by annotating a pre-existing speech corpus with an index for each phon(em)e according its prosodic environment we produce an interchangeable external source. The synthesiser then becomes a retrieval device for random-access re-sequencing that is independent of the source corpus. [Campbell, 1996].

The suitability of possible unit candidates usually was determined by means of a distance dimension. The better the candidate unit fit, the smaller the distance to a virtual ideal candidate. If the distance got bigger, the unit was assigned a penalty in terms of higher cost. The applied algorithm was trying to minimize all cost for generating a certain utterance. This way, a chain of most suitable candidate units would be concatenated to build the requested utterance. According to [Stöber et al., 1999], the assigned cost are based on two sub-dimensions. The first is the so called "unit distortion" (also "target cost" [Hunt and Black, 1996]) reflecting a penalty for the deviation of the candidate unit from certain predefined acoustic-phonetic parameters which are also known as "target specifications". Those parameters are, for example, segment duration, average fundamental frequency, phonetic context, and logarithmic intensity. The "continuity distortion" (also "concatenation cost" [Hunt and Black, 1996]), on the other hand, is defined as cost emerging from concatenating single segments. The more the candidate units deviate from each other with regard to certain acoustic-phonetic characteristics, the higher the continuity distortion. Also here, phonetic context plays an important role, next to prosodic context reflected in fundamental frequency, segment duration, and logarithmic intensity again. In contrast to target cost, continuity cost are calculated for several neighboring phones in a row. Additionally, the acoustic characteristics of the candidate units have an influence on the definition of the concatenation point. The bigger the spectral distance of probable neighboring candidates, the higher the assigned cost. [Hunt and Black, 1996] describe the selection of the candidate units as follows: All segments together define the different states of a transition network. Thus, for each state unit cost occur. Concatenation cost result from the transition to the next state. The transition network is completely connected since theoretically each phone can follow every other phone. In the event of synthesizing an utterance, the algorithm searches for the way through the network which causes the less cost. This approach shows similarities to HMM-based speech synthesis developed later on by [Yoshimura et al., 1998] and [Black et al., 2007]. The main difference is, however, that HMM-based synthesis is a probabilistic parametric synthesis technique (cf. section 3.1.1) whereas in unit selection from a natural speech corpus non-probabilistic cost functions are applied.

The Bonn Open Synthesis System BOSS

The Bonn Open Synthesis System ("BOSS") [Klabbers et al., 2001], [Breuer and Hess, 2010] deployed in the current research to implement and evaluate an independent fast speech corpus recorded to generate more natural as well as still intelligible speech at fast speaking rates is an open source synthesis software which arose from the speech synthesis architecture applied in the "Verbmobil" project for the first time [Stöber et al., 2000]. The Verbmobil synthesis system was one of the lead architectures for German speech synthesis based on multi-level non-uniform unit selection. Subsequently, the BOSS project started in the year 2000 at the "Institute for Communication Research and Phonetics" of the University of Bonn. [Stöber, 2002] designed and implemented the base system which consisted of a set of tools for the preparation of speech corpora, a utility library, and a signal processing library. The core synthesis application, called "BOSS Server", included the unit selection algorithms and a transcription module, as well as a demonstration client for text-to-speech synthesis.

After its introduction at Eurospeech 2001 ([Klabbers et al., 2001]), BOSS has been under constant development, and a number of modules and tools have been added and extended, among them classes for decision tree-based grapheme-to-phoneme conversion, duration prediction, pitch and duration manipulation, and soft concatenation of units. The focus was put on the redesign of the core architecture to provide for an easy exchange of individual modules [Breuer, 2009], [Breuer and Hess, 2010]. Also in more recent versions the BOSS architecture still separates algorithms from data as much as possible [Hammerstingl and Breuer, 2004]. It is divided into two main programs, the client and the server [Breuer and Hess, 2010]. The client contains the text pre-processing. Thus, it is the component of the architecture performing the custom-designed adaptation of the input before its results are sent to the server. The server, on the other hand, generates the synthetic speech signal and returns it to the client. For the time being, the integrated server modules perform the phonetic transcription, the duration prediction, the unit selection, and the actual speech synthesis. This structure allows for an easier adaptation to new languages and speaking styles. Also the adaptation of the duration prediction module and cost functions have been of interest in more recent research, in addition to the efficient development and preparation of new speech corpora (cf. [Breuer et al., 2006a], [Bachmann and Breuer, 2007], [Demenko et al., 2008], [Demenko et al., 2010]). The investigations conducted during the course of the present work concentrated on the development of a new corpus for a different speaking style, fast and clear speech, as well, next to inventory preparation and adaptation of duration prediction (cf. chapter 7). For further details on BOSS's architecture, modules and tools, the interested reader should refer to [Klabbers et al., 2001] and, depicting more recent developments in detail, to [Breuer and Hess, 2010].

Besides, the "BOSS-SAMPA BOSS BLF label file format" [Breuer et al.,

IPA	BLF	(X-)SAMPA
[?]	?	? (Q)
[h] / [ɦ] + vowel	single phones	h + vowel
[j] + vowel	single phones	j + vowel
[ʋ] / [v] + vowel	single phones	v + vowel
[ʀ] / [ʁ] / [ʁ̞] / [ʁ]	single phones	r + vowel
[l] + vowel	single phones	l + vowel
[ən] / [n]	@n	@n
[əm] / [m]	single phones	@m
[əl] / [l]	single phones	@l
[j] / [ʋ] / [v] / [ʀ] / [ʁ] / [ɾ] / [r] / [l] + [ən]	single phones	j / v / r / l + @n
[j] / [ʋ] / [v] / [ʀ] / [ʁ] / [ɾ] / [r] / [l] + [əm]	single phones	j / v / r / l + @m
[j] / [ʋ] / [v] / [ʀ] / [ʁ] / [ɾ] / [r] / [l] + [əl]	single phones	j / v / r / l + @l
[ts]	ts	ts
[pf]	pf	pf

Table 3.1: Unit definitions in IPA, BOSS-SAMPA (BLF) and (X-)SAMPA after [Breuer et al., 2001].

2001], designed for BOSS II, is important for the investigations conducted later on and thus examined in more detail here. The BOSS-SAMPA BOSS BLF label file format was developed to minimize inconsistencies in labeling between human labelers and, most notably, to allow for aggregation of several phones into one label since in the past it had been shown that the one-to-one attribution of one SAMPA symbol to one phone in a continuous speech signal was arguable. Especially phone sequences prone to heavy coarticulation were highly problematic and evoked inconsistencies in labeling. Thus, it was assumed that the quality of the synthesized speech could be enhanced by avoiding the usage of single phones hardly to segment as a single concatenation unit. [Breuer et al., 2001], [Breuer and Abresch, 2004], [Breuer, 2009] showed that concatenating units at very sensitive concatenation points negatively influenced the quality of the synthesized speech. They proposed the implementation of a revisited modified speech unit concept. This proposal will be reviewed and applied to fast and clear speech as outlined in chapter 8.1.

The basis for the aggregated labels used in the BOSS-SAMPA BOSS BLF label file format was (X-)SAMPA. The symbol inventory was optimized for usage in an automatic phone segmentation algorithm and, later on, for the concatenation of segmented multi-phone units. (cf. 8.1). The aggregated symbols included unaccented syllables where reduction and assimilation often are more distinct than in accented syllables (cf. chapter 2.2.3). Additionally, sequences of semi-vowels or liquid-plus-vowel as well as [h]-plus-vowel were defined as one aggregated unit. Table 3.1 shows the phone sequences aggregated to one segmental symbol (unit) in BLF format. Most of the combinations are only applied to syllable onsets. In addition, German approximants as well as those appearing in English or French loan words are affected by severe coarticulation as well, next to allophones of [r], and are therefore included here.

3.2 Modeling Speaking Rate

When implementing natural fast speech as a unit selection corpus in speech synthesis, the necessity and feasibility to adapt the duration prediction module are to be considered. Furthermore, methods and algorithms to accelerate artificially generated speech to even higher speaking rates than natural ones are of interest. Thus, in the following section different approaches of duration prediction are discussed at first. Afterwards, selected methods applied to accelerate artificially generated (fast) speech will be examined. The results of their application are described later on in chapter 7.2.2 and 8.2.

3.2.1 Duration Prediction

An adequate duration prediction enhances the perceived naturalness of synthetic speech [Brinckmann and Trouvain, 2003]. At the same time, the duration of speech segments is affected by many different factors. Several models have been developed to describe and predict the duration of speech units by considering those factors to different extents. The most common models dealing with duration prediction for speech synthesis systems will be discussed in the following section. Herein, the explanations follow the widespread distinction between rule-based and statistical approaches [Carlson, 1991], [Möbius, 2003]. Another important differentiation between models is - similar to the discussion about the optimal unit to measure speaking rate in section 2.1 - the one about the size of the unit whose duration is to predict.

As one of the first researchers dealing with duration prediction for artificial speech generation, [Klatt, 1979] assumed single phones to be the prior units to use for duration modeling. In his approach, he tried to describe and replicate perceptually important "first-order effects" of durational changes. In this connection, the rule system the researcher developed was based on the "inherent duration", the average duration specific to each phone, which he derived from accentuated phones read aloud in context by himself. Additionally, each phone was assigned a "minimal duration". While considering several contextual factors as well, he computed the phone duration to be generated by his speech synthesis-by-rule program by adding a calculated percentage of durational increase or decrease to the minimal phone duration for each phone [Klatt, 1979]:

DUR $\bar{[}$(INHDUR - MINDUR) * PRCNT]100 + MINDUR

This way, it was not possible to go below the minimal duration of each segment. Contextual factors influencing the phone duration were lexical stress, morpheme and word boundaries as well as the syntactic structure of the utterance to be synthesized. Next to the phonetic component and the actual formant synthesizer, a phonological component output a detailed phonetic and prosodic representation of the utterance to be generated, including an acoustic duration for each segment. In total, [Klatt, 1979]'s system comprised 52 segments as well as three stress markers, three types of boundary indicators, and six syntactic structure indicators serving as contextual factors which seemed to be enough to sufficiently approximate durational patterns of natural speech. The rule set consisted of eleven rules only which were developed by means of a trial-and-error method attempting to match durations occurring in natural speech utterances. In this model, speaking rate was controlled by an additional variable. It was possbile to set the rate of speechvto a value between 60 and 300 words per minute. Slow speaking rates, however, were mostly realized by inserting pauses. The developed rules accounted for 84% of the observed total variation of Klatt's speech with a standard deviation of 17 ms which was less than the "Just Noticeable Difference" for durational changes in speech as defined by [Quené, 2007]. [Klatt, 1979] himself pointed out that the order of the rules was crucial: Segmental insertion or deletion rules were to be applied preceding duration manipulation rules as otherwise duration proportions may get distorted. According to the researcher, it was also necessary to make a more detailed distinction between content words and function words since content words were more important for understanding the content of an utterance than function words were. However, as the durational rules were obtained from a single speaker's speech, the author rightly underlined that further research would be necessary to gain more precise rules more commonly applicable.

In 1985, [Elsendoorn, 1985] investigated the acceptability of temporal variations in synthetic speech for Dutch on the basis of an implementation of simple durational rules. Her results showed that already a limited set of temporal rules allowed for the generation of synthetic speech at an acceptable quality level. However, depending on the speech units used for synthesis the rule set became quite complex. The author found that changes in vowel duration were dependent from the consonant following the vowel, the number of syllables in the respective word, and if the syllable in question was accented or not. Those results were consistent with previous findings by [Nooteboom, 1972] who investigated temporal structures for Dutch as well. Another set of durational rules was implemented by [Allen et al., 1987]. The researchers developed the "MITalk duration model", hallmarked by successive duration rules which were manually adapted to special cases. However, [van Santen, 1998] and [van Santen, 1994] noted that the initial model needed significant modification to describe the interactions between other, additional prosodic factors. Besides, [van Santen, 1998] stated that he also observed many other interacting factors influencing the duration of individual segments. In connection with duration modeling taking all of those factors into account he pointed to the problem of "lopsided sparsity" which can be observed in speech: "[T]he number of rare vectors is so large that even in small text samples one is assured to encounter at least one of them." [van Santen, 1998].

Referring to the phone duration rules proposed by [Klatt, 1979], [Campbell, 1987] claimed those rules not to be sufficient to reproduce natural ("real") speech variability. Thus, he re-synthesized a given text by means of synthesis-by-rule, and compared the generated phone durations to their original durations in natural speech. One important finding was that the consonant cluster rule proposed by [Klatt, 1976] reduced the duration of affricates too much. Therefore, [Campbell, 1987] applied a simplification of durational rules designed earlier by [Crystal and House, 1986]. However, the refinement of those rules brought only little improvement compared to the previous application. Thus, in his subsequent work [Campbell, 1988b] took a closer look at the prediction of segment duration at a local as well as at a global level. The relevant factors to take into account for duration prediction at both levels were stress, phonetic and phrasal context, next to inherent differences in single segments. The author observed speaking rate related variation within texts and therefore assumed that the relative lengths of segments was also important for a correct duration prediction. Similar observations were already outlined in relation to speaking rate quantification as discussed in chapter 2.1. Also [Pfitzinger, 1996] pointed out that speaking rate needed to be defined on a local as well as on a global level. He suggested to adopt syllable based duration measurements to reflect the complex relationships of single segment durations. [Campbell, 1988b] instead proposed to apply a normalization of syllable duration while referring to the actual syllable structure. Thus, durational rules would not allow for modeling local speech rate variation in too much detail. Local speech rate was simply adding up on segment duration. In [Campbell, 1988a], "part of speech" and "position in the utterance" were added as factors to take into account in duration prediction. The researcher stated that especially the regularity of the duration of segments predicted by rule was a factor which contributed to the unnaturalness of synthesis-by-rule systems. Hence, specific additional rules were needed to cover for more fine grained variation of duration changes as well as for the timing of speech segments. In this regard, the author pointed out that "some of the durational variability of segments and syllables can be accounted for by a speech-rate-factor." [Campbell, 1988a].

Setting forth his research, also [van Santen, 1994] investigated the duration of phonetic units dependent on more numerous contextual factors than before. Following the author, to make speech synthesis sound more natural the influence of those contextual factors needed to be reproduced reliably by the duration prediction module of the speech synthesis system. Therefore, he developed a new approach consisting of a set of durational models which were based on equations of sums and products, the so called ”sums-of-products model”, capturing different types of interaction patterns. The system of durational models consisted of a category tree which included various sums-of-products models for cases behaving similar in speech production. [van Santen, 1994] explained that he preferred to neither use statistical nor durational models alone since statistical models were not able to deal with frequency imbalance whereas duration models were not applicable to reproduce factor interactions. However, he observed that factor interactions were quite regular such that most effects were intensified or diminished during interaction. Thus, all effects could be described by the developed sums-of-products model which eventually was considered as a generalization of conventional duration models.

[Kohler, 1988] introduced a modification of [Klatt, 1979]'s rules for German. He realized that there was a difference between universal versus language-specific durational rules. The number of rules for minimum phone durations was reduced to describe phoneme classes, not single phones anymore. After those minor adaptations, the approach of rule-based duration prediction as proposed by [Klatt, 1979] seemed to be applicable to German as well [Möbius and van Santen, 1996] developed a model of segmental duration in German as well. Their duration system made use of the ”sums-of-products” model proposed by [van Santen, 1994]. They took into account various textual, prosodic, and segmental contexts. The duration of each phone was predicted depending on its feature vector. Next to the fundamental steps of setting up the category tree and defining the specific sums-of-product model for each leaf, it was important that factors were computable from the text to be applicable in a TTS system.

In 1991, [Campbell and Isard, 1991] introduced an approach to duration prediction based on the syllable as underlying speech unit. The ”elasticity hypothesis” derived from this concept stated that the single constituents of a syllable adapted their relative duration to the durational frame of the entire syllable. Herein, the durational frame of a syllable is their measured or given duration. Factors influencing syllable duration and taken into account in this duration prediction model were ([Campbell and Isard, 1991]: 38f):

- The number of phonemes in a broad transcription of the syllable.
- The nature of the syllabic peak; whether it is a tense or lax vowel, a diphthong, or a sonorant consonant.
- The position of the syllable in the foot.
- The stress assigned to the syllable, and the nature of any pitch movement associated with the syllable.
- The function/content role of its parent word.

When the ”strong” form of the elasticity hypothesis was reviewed, it occurred that neither an additive-linear nor a multiplicative-linear relationship existed between the durational changes each segment of a syllable underwent while adapting its duration to the syllable frame. Therefore, weaker forms of the elasticity hypothesis were developed to take into account additional factors, such as position of the syllable within the respective utterance, the position of the segment within the respective syllable, or the influence of differing phonetic contexts. For short syllables, investigations showed that the expansion or compression was the same for all syllable constituents. This was in line with the initial hypothesis that within a given syllable a constant factor was to be applied for the shortening or expansion of the respective syllable segments. However, for long syllables a significant difference was found for the mean values of consonants in syllable onset and coda, as well as for the mean values of vowels contained in the nucleus versus the coda. The evaluation of syllables of average duration showed similar results. After further investigation, the authors had to admit that indeed the strong form of the elasticity hypothesis was not generally applicable and needed refinement. Nevertheless, the results they revealed were similarly accurate as those accomplished by the segmental duration prediction model by [Klatt, 1979]. [Campbell and Isard, 1991] concluded that timing processes in speech seemed to take place on three different linguistic levels: On the utterance level where boundary effects were of main importance, on syllable level where stress and rhythm had the most influence, and on segmental level where phonetically motivated effects were of main interest.

During their research, [Portele et al., 1994], [Kraft and Portele, 1995] and [Meyer et al., 1995] elaborately discussed phone-based versus syllable-based duration prediction. A perceptual evaluation of utterances generated with different underlying duration prediction models showed significantly better results for the syllable-based method applied, and beyond that was more easy to implement. The authors found the factors influencing the duration of the different speech units to be the same as already noted previously. In their subsequent investigations about duration prediction as part of the prosody generation in speech synthesis, [Meyer et al., 1995] revealed that syllable-based duration prediction still performed better than phone-based duration prediction but was not perceived as being as good as natural durations which was consistent with the findings of [Brinckmann and Trouvain, 2003] discussed previously. In line with [Möbius and van Santen, 1996], also here it was pointed out that the position in the phrase, the stress level, the syllable structure and the number of syllable segments were the most important factors influencing segment duration.

More recently, a different method of duration prediction was developed and successfully deployed in speech synthesis: the so called CART-based approach (Classification And Regression Trees, [Riley, 1990]). Classification and regression trees are derived from a more general purpose statistical method, and include a statistical learning algorithm. A CART is constructed by making binary decisions on given factors, for example acoustic and/or segmental characteristics, to minimize the variance of the duration of the resulting subsets. [van Santen, 1994] pointed out that CART-based methods cannot make use of a given ordered structure; this had to be developed during statistical processing. [Riley, 1990] chose as factors (”features”) influencing the segmental duration phonetic context, stress level, and lexical position, next to the identity of the segment itself. Generally speaking, a careful choice of the feature set is always necessary. The CART algorithm statistically selects the most relevant features and tries to find the classification of features that gives the minimum error. Categorical features are permitted to result in ”classification”, and continuous features to result in ”regression”. When applying CART to his data, [Riley, 1990] found that segmental duration was influenced by both categorical and continuous features. Proceeding in his discussions, the author stresses the advantage of the algorithm giving ”honest” (in terms of ”objective”) estimates and enabling human interpretation. It can also be applied to detect an end-of-sentence. An investigation by [Brinckmann and Trouvain, 2003] showed that the performance of CART in terms of accuracy of predicted durations compared to measured durations in natural speech and analyzed in terms of error rate and root mean square error (”RMSE”) was significantly better than duration prediction based on [Klatt, 1979]’s rules. Moreover, CART was also preferred in a perceptual evaluation, presumably as the segmental durations generated were closer to natural speech segment durations. To conclude, this method will be applied in the current research to predict segmental duration in normal and fast and clear speech. Results will be compared to evaluate whether performance is equally good for both speaking styles. CART application and the analysis of its results are outlined in section 7.2.2. However, although CART may generate more accurate predictions (cf. [Breuer et al., 2006b], [Klessa et al., 2007]), at the same time it may also show more variability in its predictions. For insufficient or sparse training data, predictions showed to be rather poor (cf. [van Santen, 1994]). Nevertheless, CART is seen as a promising approach to segmental duration prediction as no hand-crafting of durational rules is necessary and large datasets can be handled easily. For this reason, CART-based duration prediction is also applied for the adapted duration prediction for fast and clear speech as described in chapter 7.2.2.

Considering the observation that several temporal units are to be taken into account for more accurate duration prediction (cf. [Keller and Zellner, 1996], [Breuer et al., 2006b]), [Campbell, 1990] proposed to use ”neural networks” to predict syllable duration instead of conventional approaches, because it would allow for the representation of an even larger number of factors as well as the complexity of their interaction. So called ”feature vectors” would be the result, including nine levels of description for any feature. The training corpus of spontaneous natural speech the author used to train the neural networks revealed a large variability of timing, and thus included some amount of uncertainty. However, a respective weighting of factors in the neural network would lead to a reduction of the mean square error. Thus, the best average duration would be estimated after sufficient training. Experiments showed that the number of phonemes contained in the syllable was mostly highly correlated with the syllable duration itself. The time needed to apply the neural network was negligible compared to that needed for rule-set implementation; however, further perceptual testing would be required to validate the author's findings. Therefore, neural networks are not applied in the current research.

3.2.2 Speech Rate Modeling

In addition to the accurate prediction of segmental duration to enhance naturalness and intelligibility of synthesized speech, the modeling of speaking rate is another processing step of particular interest when investigating implications of fast speech implementation in speech synthesis. Especially when it comes to the production of speaking rates beyond natural speaking rate, additional modification of the generated speech signal is required. Therefore, in the following section different speech acceleration and time compression techniques are discussed.

Modifying speaking rate is quite easy in parametric synthesis approaches. It can be done by simply increasing or decreasing the factors determining segment duration. All other characteristics of sounds to be generated are kept; therefore, synthetic speech generated with parametric synthesis often still includes specific characteristics of the speech sounds important for correct identification and perception, such as transitions and certain coarticulatory phenomena (cf. chapter 4.2). However, in concatenative speech synthesis this procedure is not applicable since natural speech units are used to generate the desired utterances.

Linear Time Scaling

A classical and thus quite common approach to modify prosody in concatenative speech synthesis is the use of pitch synchronous overlap add techniques, *PSOLA* [Charpentier and Stella, 1986]. It allows for the modification of prosodic parameters, especially speaking rate and intonation, while keeping a high level of naturalness of the manipulated speech. PSOLA can be applied either as FD-PSOLA (frequency domain) or TD-PSOLA (time domain). The application of the algorithm requires a preliminary pitch period labeling of the input waveforms. While applied, it simultaneously controls the value of the synthesized pitch and the duration of the synthesized signal in three steps: First, the speech waveform is analyzed by means of an analysis window in order to produce a preliminary short-time representation of the speech signal. Second, the required modifications are applied to this preliminary speech signal representation. The speech waveform is multiplied by a sequence of time-translated analysis windows centered around pitch marks; their length is proportional to the local pitch period but slightly longer than one period producing a slide overlap. As a third step, the modified speech signal gets generated. Herein, the synthesis process consists of a mapping between synthesis time instants and analysis time instants according to desired prosodic modifications [Syrdal et al., 1998]. In the most simple case, the resulting signal is a simple linear combination of the analyzed and translated versions of the original signal. An acceleration of the speech signal can be achieved by applying TD-PSOLA and selectively eliminating some of the short-term analysis signals. The critical limit for speech acceleration is factor two; for factors above two, a tonal noise is introduced for the unvoiced signal parts which can be avoided by additionally using FD-PSOLA. According to [Moulines and Charpentier, 1990], TD-PSOLA is computationally very efficient and therefore recommended to use for speech rate acceleration in concatenative speech synthesis. On the other hand, due to limited smoothing and prosodic modification possibilities it is less feasible for non-parametric synthesis approaches. TD-PSOLA is outperformed by "Harmonic plus Noise Models" with regard to intelligibility, naturalness and pleasantness; only with regard to computational load, it is much better.

An alternative method to linear acceleration is non-linear time scaling. Non-linear time scaling approaches often include mimicking characteristics of natural fast speech with regard to pause duration, intonation, and prosodic breaks. Thus, they could be considered as an alternative to PSOLA since they are aiming for replicating the more natural, non-linear acceleration patters of natural fast speech without comprising undesirable phenomena like coarticulation and reduction. [Trouvain, 2002a], [Trouvain, 2002b], for example, suggested to control for speech tempo in speech synthesis by prosodic phrasing. However, the proposed model was restricted to prosodic phrase breaks with implications for pausing and phrase-final lengthening. To speed up, predicted prosodic breaks were skipped and less phrase-final lengthening was applied. However, the author observed that this approach was not applicable to produce fast speech, mostly reflected in inconsistent listener ratings. Moreover, it was not possible to generate ultra-fast speaking rates. Other methods of non-linear compression aim at manipulation on segmental level. [Covell et al., 1998], for example, proposed an algorithm called "MACH1" for non-linear compression of speech. The algorithm compressed the components of an utterance to a different degree to reproduce the timing of natural fast speech based on phoneme classes and stress levels. The researchers observed a drop of intelligibility of speech linearly accelerated to a speaking rate of 270 words per minute or more whereas for non-linear acceleration word intelligibility only decreased at a speaking rate of 500 words per minute. The difference in comprehension rate between "MACH1"-compressed and linearly compressed speech increased with increasing compression rate. [Covell et al., 1998] concluded that "MACH1" offered significant improvements in comprehension compared to conventional linear compression techniques, especially in short dialogs at high compression rates.

An algorithm called "WSOLA" was developed by [Demol et al., 2005]. Speech segments were assigned one of five acoustic classes, based on signal energy and an "Average Magnitude Difference Function". Depending on the assigned class, segments were time-scaled with different factors. Additionally, WSOLA added a tolerance interval to the original overlap add approach to ensure signal continuity at segment joints. To evaluate the generated fast speech, two different versions of time-scale coefficients were applied: Speeding up consonants more than vowels and speeding up vowels more than consonants to reduce redundant information. The latter was in accordance with natural fast speech production. Pauses were sped up most, plosives on the contrary were not time-scaled at all. Both methods were not judged significantly different from natural fast speech in a listening evaluation, but results revealed that speeding up with vowel-like segments being faster than consonants was slightly preferred, even to natural fast speech. However, the factors chosen for time-scaling were within the range of natural speech tempo. The authors assumed that when exceeding this range preferences may have become clearer. They concluded that the best strategy seemed to be to leave a maximum of information intact, especially in consonant-like segments, and speed up vowel-like segments more because they contained more redundant information. [Höpfner, 2008] proposed a similar approach also for German. He observed that the MACH1-algorithm suggested by [Covell et al., 1998] did not reveal satisfying results as especially the strong consonantal shortening decreased the intelligibility of the generated speech. Instead, the researcher suggested to pay special attention to plosives and fully keep them, in line with [Demol et al., 2005]. Applying a double loop TD-PSOLA, the algorithm [Höpfner, 2008] proposed generated speech with higher intelligibility for a speaking rate of twelve syllables per second than speech manipulated by means of WSOLA and TD-PSOLA. However, correct segmentation showed to be problematic, and wrong segmentation led to phone confusion and misunderstanding.

In his research about tempo changes in speech production and its implications for speech synthesis, [Trouvain, 2004] conducted several perception experiments where he mainly manipulated pause duration in a diphone TTS system to slow down the speaking rate, depending on syntactic and prosodic characteristics. The researcher found that very slow speech generated this way was preferred over slow speech linearly slowed down from normal speech. However, for "medium slow" speech this preference was not confirmed (cf. also [Moos and Trouvain, 2007]). Consequently, [Trouvain, 2004] introduced a new durational model where number and duration of pauses as well as the factor for phrase-final lengthening were adapted accordingly. Anyway, this approach is non-transferable to the artificial production of accelerated, fast speech in unit selection speech synthesis. Since [van Santen, 1997] had pointed out that changes in speaking rate were not uniform, [Janse, 2001] conducted a series of experiments comparing word-level intelligibility after linear acceleration with word-level intelligibility after non-linear time compression. She found that the intelligibility of words linearly compressed was rated higher than the intelligibility of non-linearly compressed words. Setting forth her research, [Janse, 2002], [Janse, 2003b], [Janse, 2004] made the temporal patterns of artificially time-compressed speech more similar to that of natural fast speech. The results of another series of perception experiments conducted showed that mimicking natural fast speech phenomena did not improve intelligibility over linear time compression. Further details of this for the current research important series of investigations are discussed in 4.2.2. Based on these findings, [Janse et al., 2007] eventually concluded that, although PSOLA manipulation resulted in unnatural speech per se - since natural speech acceleration is non-linear - the intelligibility of speech accelerated linearly by applying PSOLA was always judged better (more intelligible) than speech accelerated based on natural acceleration patterns.

Thus, despite the promising results for the application of non-linear time scaling methods outlined above, the observations made by [Janse, 2002], [Janse, 2003b], [Janse, 2004] led to the conclusion that linear acceleration by means of PSOLA was better applicable to generate fast and clear speech stimuli for the series of perception experiments conducted in the frame of the work presented here (cf. chapters 7.1.1, 8.1.3, 8.2.1). Moreover, the complexity of non-linear acceleration models - which for fast speech might even be higher than for normal rate speech - was another reason for choosing linear time compression for the acceleration of the speech stimuli generated for the current research. Moreover, TD-PSOLA was seen as more mature and most commonly used to accelerate speech in speech synthesis, so no additional unpredictable factor was introduced into analyses. However, the application of non-linear time scaling approaches might be a field of further research to generate faster speech with higher intelligibility.

3.3 Summary and Conclusions

Since the synthesis technique used plays a crucial role depending on the goal of research, different synthesis techniques and their advantages and disadvantages were discussed in this chapter. It was explained why concatenative unit selection synthesis is the first choice if the generation of natural sounding speech is the target of investigations. However, smooth transitions required by the emergence of coarticulatory phenomena during articulation processes are crucial for the intelligibility of natural as well as synthetic speech; those are modeled best through parametric speech synthesis. For the current research, the differences between parametric synthesis, represented by the commonly used ”JAWS Eloquence” application [FreedomScientific, 2011], and the concatenative unit selection synthesis system ”BOSS” [Klabbers et al., 2001] are of main interest. BOSS's system architecture was detailed in section 3.1.2. The perceptual evaluation of fast speech generated at different speaking rates with different speech synthesis systems is the major investigation of the work presented here. It is outlined in detail in chapter 8.2.

Afterwards, different duration prediction methods were discussed. An adequate duration prediction enhances the perceived naturalness of synthetic speech [Brinckmann and Trouvain, 2003]. In the course of time, numerous models have been developed to describe and predict the duration of speech units by taking into account various factors to different extents. Since the duration of speech segments is affected by many different factors, the implementation of natural fast speech as a unit selection corpus in speech synthesis most presumably requires an adaptation of the duration prediction module. To this end, the feasibility of doing so will be examined by applying the most common and promising approach, classification and regression trees (CART, [Riley, 1990]). CART is to be seen as a promising approach to segmental duration prediction as no hand-crafting of durational rules is necessary and large datasets can be handled easily. CART-based duration prediction was applied to both the normal and the fast and clear speech corpus. A comparison of the most important features applied for CART building is outlined in chapter 7.2.2. The results will lead to the conclusion whether or not CART-based duration prediction is also applicable to predict segmental duration in fast and clear speech.

Subsequently, possibilities to model speaking rate in a TTS system were described. Although it has some known disadvantages, especially the introduction of noise for an acceleration factor of two or higher, pitch-synchronous overlap add (PSOLA) is the most commonly used algorithm for such a task. [Janse, 2003b] showed that speech linearly compressed by means of PSOLA was judged significantly better than fast speech generated by mimicking natural fast speech patterns. Therefore, PSOLA was chosen as the algorithm to use for the generation of fast speech, be it natural or synthesized, at speaking rates higher than any natural speaking rate. Investigations on accelerating natural speech in normal as well as fast and clear speech tempo to higher speaking rates by means of PSOLA are outlined in section 7.1.1. Experiments with speech synthesized at fast speaking rates with different underlying speech synthesis systems and accelerated by means of PSOLA to even higher - and therefore highly unnatural - speaking rates are presented in chapter 8.2.

Chapter 4

Fast Speech Perception

When implementing fast speech as a separate speaking style in a speech synthesis system, also its perception needs to be taken into consideration. Therefore, the following chapter discusses several important aspects of speech perception. First, the perception of natural fast speech is described in section 4.1. Common models of speech perception as well as investigations about units of speaking rate perception are examined. Afterwards, the perception of artificial fast speech is described in chapter 4.2. Evaluation methods for artificially produced fast speech in general are presented in section 4.2.1. Subsequently, the perception of time-compressed natural speech is discussed in section 4.2.2. Implications for the evaluation of the perception of synthesized fast speech are finally outlined in section 4.2.3.

4.1 Natural Fast Speech Perception

Along the lines of the explanations regarding speaking rate production and quantification given in chapter 2.1, common models describing speaking rate perception are examined at first in the following section. Mechanisms of "perceptional adjustment" and compensation with regard to durational as well as spectral characteristics of fast speech are described. Afterwards, the perception of speaking rate as such is explained in chapter 4.1.2. Possible units of speaking rate perception are discussed.

4.1.1 Models of Speech Perception

A detailed overview over well-established theories and approaches about speech perception is given by [Diehl et al., 2004]. In their explanations, the authors outline three main theoretical perspectives on speech perception in general. The first one is the *Motor Theory* of speech perception (cf. [Liberman, 1996]) which claims that the perceived phonemes and features have a more simple relation to articulatory events compared to acoustic or auditory events. The underlying references are the "intended gestures" rather than more peripheral acoustic events. This assumption is based on the hypothesis that the objects of speech perception must be more or less invariant. The complex mapping between phonemes and their acoustic realizations are mainly attributed to coarticulation. In this perspective, the human ability to perceive speech depends on a specialized decoder or module which is seen as speech-specific, unique to humans, and innately organized. [Liberman, 1981] had claimed already in 1981 that ”speech [was] special” when processed by human listeners. This view is also reflected in the observations made by [Lehiste, 1994] (after [Wagner, 2005]): The author found that the listeners' sensitivity for variations in duration, F0 or intensity differed significantly between speech and non-speech stimuli. Another approach discussed by [Diehl et al., 2004] is the *Direct Realist Theory*. Like in the Motor Theory approach, the objects of speech perception are articulatory rather than acoustic events, but in contrast to Motor Theory, the Direct Realist Theory considers that objects of perception are the actual vocal tract movements. Additionally, it denies any specialized mechanism being applied in human speech perception. Structuring the speech signal, however, plays a vital role: Gestures are seen as being co-produced, but remaining separate and independent from each other.

The third group of theories considered are so-called *General Auditory and Learning Approaches*. According to [Diehl et al., 2004], they do not include any special mechanisms in perception either but rather advocate more general mechanisms in perceptual learning for any environmental sound. The listeners' recovery of spoken messages is neither equivalent to nor mediated by the perception of gestures. The general claims of theories belonging to this group of approaches comprise categorical perception and phonetic context effects, for example the so called ”stimulus length effect” where changes in transition duration as well as the durational contrast to neighboring segments play an important role, next to the compensation for coarticulation, for example for context sensitive acoustics. This is in contrast to the above mentioned approaches of Motor Theory and Direct Realist Theory which assume that the intended gestures are invariant even though the acoustics are variable. From this point of view, effects of coarticulation serve as information for the identity of the context segment as opposed to information masking the identity of the target segment. Moreover, General Auditory and Learning approaches state that the listener needs to learn to distinguish between speech and non-speech categories: What is important, what is not important for communication in a specific language? The gathered experience with speech sounds leads to the creation of category prototypes (”prototype effect”) or high-density representations of exemplars that act as ”perceptual magnets” (cf. [Kuhl, 1991]). A very common statement of these approaches is ”production follows perception, and perception follows production.” Herein, the sound systems of languages tend to satisfy the ”Principle of Dispersion”, whereby inter-phoneme distances are maximized within the available phonetic space to promote the intelligibility of utterances. The ”Auditory Enhancement Hypothesis”, for example, formulates the implementation of this dispersion principle in a short sentence: "Create maximally distinctive sounds by precise articulation." This principle is in line with certain speaking strategies discussed in chapter 2.3 and required for the fast and clear speech corpus recordings (cf. section 7.1) to be implemented in a unit selection speech synthesis system.

Taking the different models of speech perception into account, [Anward and Lindblom, 1997] noted that no agreement could be found in the literature on the mechanisms which made speech perception so singularly fast and robust. Gestural accounts like the Motor Theory which hypothesize a faster phonetic module for speech contrast with the Direct Realism approach where no such module is required, but an extraction of perceptual invariants takes place. Moreover, both approaches are in opposition to theories which advocate speech being structured by auditory or acoustic goals, as for example by compensation for coarticulation. Acoustic constancy is attributed to speech production being under output- or listener-oriented control (cf. chapter 2.3). Instead, [Anward and Lindblom, 1997] pointed out that according to their view *exemplar-based models* of speech perception as a type of mechanism whose prospects of meeting the criterion of processing speed appeared particularly favorable are the most promising. In such models, phonetic and grammatical entities and rules arise developmentally as emergent consequences of the listener's cumulative perceptual experience. Furthermore, exemplar models assume that memories, including phonetic ones, are built by experience without speaker normalization (cf. [Johnson, 1997], [Pierrehumbert, 2000]). According to [Anward and Lindblom, 1997], the assumption is that a category is defined by all perceived instances of that category, and their auditory attributes are read into a phonetic memory again every time an instance is perceived. The result of this process is a network of sound-based memory structures each one linking sound to meaning. Exemplar clouds are built by judging the similarity of a stimulus to other instances that had been perceived previously. Coarticulation does not jeopardize the separability of categories, since distinctiveness is more important than any kind of invariance.

Following this approach, [Wade and Möbius, 2007] created a model of speech perception which did not consider speaking rate nor lower level temporal cues explicitly. Instead, the authors proposed that newly encountered speech signals were encoded as sequences of detailed acoustic events specified in real time at salient landmarks and compared directly with previously heard patterns. According to [Wade and Möbius, 2007], their model performed similarly to human beings in relying on temporal information for consonant and vowel recognition. Experimental results indicated that compensation for speaking rate in human perception may follow implicitly from even modest knowledge of robust correlations between temporal and other properties of individual speech events and those of their surrounding contexts, and do not require special normalization processes. However, increasing rate variability was found to negatively affect the accuracy of perception which caused an increased cost of encoding rate variability in memory. Therefore, the researchers suggested a simpler mechanism of perception: A pure acoustic exemplar approach to representation and comparison. It assumes that the memory contains an ordered collection of richly specified, real-time acoustic descriptions of previously perceived sounds in different collections. Perception then involves comparing a newly encountered acoustic signal in space and time with the entire memory, and identifying feature labels occurring near regions of maximum similarity. Acoustic descriptions would take the form of potentially informative parameter values extracted at salient landmark locations in the signal. Temporal information would be encoded in the locations of acoustic landmarks with respect to each other in time.

If objects of perception were invariant, however, they would require compensation for deviation and normalization of realized speech characteristics. Since both durational and spectral characteristics of speech are heavily influenced by changes of speaking rate as discussed previously in section 2.2, a closer look must be taken at perceptional adjustment and compensation mechanisms presumably playing a vital role in fast speech perception. Thus, two different approaches to perceptual normalization are examined in the follwoing (cf. also [Wrede, 2002]). The one is about so called ”intrinsic timing”. Intrinsic timing approaches assume that certain characteristics of speech are independent from the actual speaking rate. Such characteristics can be observed when assuming ”durational normalization” to take place during perception. On the contrary, ”extrinsic timing” theories hypothesize that spectral characteristics are normalized after a rough speaking rate analysis took place to define how to normalize single acoustic characteristics (”spectral normalization”). [Wayland et al., 1994] assume that both intrinsic and extrinsic timing occur in parallel, the former on syllable level, the latter on a more global level because the perception of speaking rate as such must take place over a longer stretch of time (cf. [Kato et al., 1997]).

Taking a closer look at the duration of perceptional units it becomes obvious that they play an important role in speech rate perception [Lehiste, 1994]. As outlined in chapter 2.1, from an acoustic point of view the physical dimension of duration can be measured and specified by certain numerical values. However, from a perceptional point of view a perceived stimulus triggers a certain *impression*. In psycho-acoustics, such percepts are categorized into three classes: First, there are quantitative percepts which can be described by statements like ”half as loud” or ”twice as much”. Second are categorical events. Those are coupled to a quantitative change of an acoustic stimulus. The stimulus is assigned a specific phonetic category, for example a phoneme or phoneme class. The third group is called qualitative percepts. Qualitative percepts reflect a subjective impression, and are thus not feasible for (objective) experimental evaluation. An auditory event is composed of each of those percepts. The most common auditory events are subjective duration, referring to the duration of the event, loudness, related to the measurable sound pressure level, pitch, related to frequency and intonation, and timbre, related to the spectrum of a sound [Blauert, 1983]. Human beings are able to perceive auditory events as single percepts or as a combination of those. [Zwicker, 1982] was able to show that measurable physical duration and perceived duration were not generally concurrent. He defined the measurement *dura* where 1 dura corresponded to the subjective duration assigned to a tone with 1 kHz frequency, 1 second duration, and 60 dB sound pressure level. That way, the author revealed that the relation between physical duration and perceived duration decreased proportionally till a duration of 30 ms. Nonetheless, as soon as this critical value was exceeded further shortenings were perceived as less short or even not perceived at all. [Carlson et al., 1979] noted that in general it was not clear how precise the specification of duration was in the speech code common to speaker and listener. However, the researchers found that the sensitivity to durational changes was greater in vowels than in consonants. Moreover, they noted that the durational balance between syllable nuclei and intervals between stressed vowels were perceptionally significant (cf. [Carlson and Granström, 1975], [Huggins, 1972b], [Lehiste, 1977]).

Another aspect related to durational characteristics of speech perception is the one of the so called ”Just Noticeable Difference” (henceforth ”JND”). Already in 1972, [Huggins, 1972a] investigated the JND for segment duration in natural speech. The researcher evaluated the perceived duration of [p], [ʃ], [m], [l], and [ɔ] in different positions and contexts. Results showed that subjects were much more sensitive to changes in vowel duration than to changes in consonant duration. Moreover, changes in stress and rhythm were perceived as well. Directing subjects' attention played an important role: When listeners were asked to focus on stress, changes in duration were observed more precisely. What was perceived as an ”acceptable duration” was largely context-dependent. [Huggins, 1972a] concluded that JND may be similar for segments that were produced by means of similar articulation. Actual values measured were 40 ms for [m] and [p], and 2%-3% of the overall duration for vowels. In their investigation of the JND of articulation rate at sentence level, [Eefting and Rietveld, 1989] noted that based on their analysis the JND could be estimated to be 4.43% of the speech tempo of the standard speaking rate. However, the authors found that in a paired comparison task two factors affected the tempo judgments: First, the response category to be used by the subject (”higher tempo” versus ”lower tempo”), and second, whether the position of the stimulus was first or second in the pair formed together with the standard tempo. These findings are in line with the results [Quené, 2007] found when investigating the JND for accelerated or decelerated human speech. For accelerated tempo, the JND added up to -3% to -5%, and to +5% to +6% for decelerated tempo relative to the fundamental rate, and depending on experimental design. Interestingly, professional speakers produced a variation of up to 4% depending on the degree of novelty of the information in the relevant utterance. Tempo changes which were above the JND threshold were interpreted as being relevant for communication. The researcher concluded that a speaker may express the relevance of an utterance in a greater context by changing the tempo, and listeners could interpret a change of speaking tempo as a sign for the importance of what was currently said.

Since many phonetic contrasts are based on very fine-grained differences in duration, the question about how listeners still distinguish between categories when speaking rate increases is important. Perception experiments have shown that such differences are processed dependent on the speaking rate, and criteria for judging acoustic cues are altered by the perceptual system in relation to speaking rate [Dupoux and Green, 1996]. An investigation by [Miller and Liberman, 1979], for example, who played stimuli with transition durations of different length for [b] and [w] to their subjects, revealed that stimuli with short initial transitions were always categorized as beginning with [b] whereas syllables with a longer transition duration were categorized as starting with [w]. Moreover, also the overall syllable duration was changed during the experiment. Results showed that for longer overall syllable durations also a longer transition duration was necessary to elicit correct distinction between [ba] and [wa]. [Miller and Liberman, 1979] concluded that differences between perceptional categories were sustained when speaking rate was changed to enable listeners to correctly identify what was said. The importance of the preservation of the characteristic transitions of vowel formant frequencies to and from surrounding consonants as well as other important acoustic cues especially in fast speech to ensure the correct perception and identification of phonemes, syllables or words by potential listeners was also highlighted by [Marslen-Wilson and Tyler, 1980], [Amerman and Parnell, 1981], [Greisbach, 1991] and [Martinez et al., 1997] (cf. chapter 2.2.2). However, such fine-grained analysis would go beyond the scope of the work presented here. Thus, the fast and clear speech produced by the selected speaker was only evaluated with regard to characteristics of vocalic segments as well as overall spectral similarity to their normal speech (cf. section 6.2.1).

The spectral characteristics of vowels and consonants are severely influenced as well when speaking rate increases. Here again the question arises how a listener adapts to these acoustic variations. [Gottfried et al., 1990] evaluated the effect of speaking rate on the perception of vowels. The researchers found out that listeners adjusted their judgment and identification of vowels to the sentence rate when either duration, formant frequencies, or both were varied systematically. Especially vowels which were differentiated in natural speech by both temporal and spectral information were obligatorily identified by duration in relation to the speaking rate of the overall sentence. Similar results were revealed by [Widera and Portele, 1999]. Moreover, the authors observed that listeners also compensated for speaker-characteristic variations of vowel reduction. In a follow-up study, [Widera, 2000] detected that for different vowels different levels of reduction were perceived: Three reduction levels for [aː], four for [iː], and five for [uː]. The agreement between subjects was above 70%. However, in line with [Gottfried et al., 1990], the reduction levels were mainly influenced by vowel duration. [van Bergem, 1995] noted that if a listener expected vowels to be reduced, like in conversational or fast speech, this expectation was helpful in identifying words correctly.

Although listeners are in the position to correctly identify reduced vowels in fast speech, they are not able to do so without any additional information contained in the acoustic context of the utterance [Verbrugge et al., 1976]. An example given was the observation that a syllable containing a tense vowel cut from fast speech was judged as comprising a lax vowel instead when presented in isolation. When the syllable was put into a slow carrier sentence, the perception of the lax version of the vowel was even more persistent. This showed that the context information about the speaking rate of an utterance was used by listeners to identify vowel quality and duration. On the other hand, when a slowly produced syllable was presented in a fast spoken carrier phrase, performance did not decrease [Verbrugge and Shankweiler, 1977]. Also, identification was not dependent on the specific context in this case. The authors explained their observation with the assumption that a slowly spoken syllable inherently contained enough information about the speaking rate it was originally realized with. Thus, it could not be perceived as fast. These observations are good examples for extrinsic timing approaches since those assume that speaking rate must be identified first, before spectral characteristics are normalized.

Due to the high variability of speaking rate observable in natural (fast) speech (cf. [van Santen, 1994]), the speaking rate of the speech corpora recorded in chapter 7.1 as well as of the stimuli evaluated in chapter 8.2 were not expected to be completely consistent with regard to measurable speaking rate. However, since the JND for human speech is quite small, steps of speaking rate differences between different groups of stimuli were defined clearly outside the just noticeable range to avoid elusivemess and gain more clear and valid results from the perceptual evaluation.

4.1.2 Perception of Speaking Rate

The perception of speaking rate is mainly related to vowel duration, frequency and intensity [Carlson and Granström, 1975], [Bond and Feldstein, 1982]. [Bond and Feldstein, 1982] made the observation that with increasing vowel frequency stimuli were perceived as shorter in duration. Furthermore, the authors noted that pitch, loudness and speaking rate co-varied in natural speech. However, relations were not perfectly predictable. Since the researchers did not ask their subjects to evaluate other dimensions, they attributed all detected and rated changes in speech to changes in speaking rate. Moreover, the researchers hypothesized that repeated practicing of encoding and decoding of such speech characteristics produced a data structure in memory that incorporated not only single, specific characteristics but also interrelationships between them. In contrast, [Koreman, 2003] assumed that the overall phone rate as well as the phone deletion rate were the main cues for speech rate perception. In his investigations of articulation rate perception he noted that both underlying and surface structure of a representation had to be taken into account when speaking rate was to be modeled. Referring to [Dankovičová, 1999], the author stated that intonation phrases were the domain of speaking rate variation. A lower realized speaking rate as well as a lower intended speaking rate (cf. chapter 2.1) caused an utterance to be perceived as slower. This effect was observed to be larger the more both speaking rates differed from each other. However, although sloppy speech (r̄ealized speaking rate) contained 35% to 40% deletions the results indicated that the stronger system-oriented constraints in fast hyperspeech did not lead to a different judgment of speaking rate for clear speech uttered at a normal speaking rate. Later on, [Koreman, 2006] noted that neither the intended nor the realized phone rate as such was appropriate to explain the perceived speaking rate for each speaker subject. He hypothesized that other factors might influence the perception of speaking rate as well. Those factors were presumably dependent on pausing and/or disfluencies. The researcher concluded that he was not able to find any direct relation between articulation rate and perceived rate.

The listeners' ability to discriminate different speaking rates was also discussed elaborately in connection with the perception of speaking rate in foreign languages. While [Osser and Peng, 1964], [Vaane, 1982] and [Roach, 1998] initially argued that speech of an unknown language often was perceived as being produced at a faster rate than the own native language, already [Vaane, 1982] questioned whether this assumption was indeed correct: The results of her research indicated that the listeners' main cue for tempo detection was not their knowledge of the lexical information of the utterance, but rather some temporal features in the acoustic speech signal. In later experiments about the perception of the intended speech rate of utterances produced at different speaking rates in different languages, [Dellwo et al., 2006] indeed found that listeners were well able to identify the intended speech rate of an utterance across different languages. Additionally, they revealed an almost linear relationship between the perceived speech rate and the laboratory measurable speech rate. Using stimuli derived from the "BonnTempo corpus" ([Dellwo et al., 2004]), the authors collected listener ratings for stimuli which were produced at different intended speaking rates in the native language of the listeners as well as in two other languages. Listeners' self-estimated knowledge of the respective other (foreign) language was at a medium range. [Dellwo et al., 2004] found that for their native language listeners' agreement on speaking rate was higher than for foreign languages. Indeed, the vowel rate was an important cue for speech rate perception, which was in line with [Bond and Feldstein, 1982]. However, also for foreign languages subjects revealed some notion of what a canonical normal speech rate was like. Furthermore, they were also able to discover the intended speech rate. Nevertheless, the question on how exactly listeners identified speech rates across different languages could not be answered.

Units of Speaking Rate Perception

Similar to the problem of defining units of speech rate production (cf. chapter 2.1), also units of speech rate perception are subject to ongoing discussions. In 1998, [Kegel, 1998] still assumed that human perception included a segmentation of the perceived speech signal into smaller units, presumably phonemes, in analogy to written language where letters are the smallest units of perception. Those units were then identified by their function and meaning, and thus enabled the listener to understand the whole utterance. This assumption was in line with many previous studies on speech perception who supposed that the speech signal consisted of a sequence of separate phones, influencing each other only to a certain degree. Nonetheless, other investigations showed that this was not correct. [Greenberg, 1996], for example, noted that due to manifold overlap and coarticulation in speech it was not possible to separate an utterance into single segments. Moreover, the author also observed that understanding speech in general did not require a detailed spectral portraiture of the signal. The primary carrier of information was the temporal evolution of approximate spectral patterns whereas fine spectral detail was not required to satisfactorily reproduce speech. [Greenberg, 1996] ascribed this finding to the redundant nature of speech. To enhance Automatic Speech Recognition (”ASR”) performance, the researcher suggested to encode only a sparse representation of the speech signal including the relevant linguistic information. However, the open question then was what the minimum amount of information required was. Since [Greenberg, 1996] found that most coarticulation effects took place within the syllable, he suggested to rather assume a segmentation of the speech signal into syllables than into phones to apply in ASR. In addition, he observed that syllabic units were generally preserved in fast speech, even if a deletion of single constituents took place. To conclude with, the author noted that the speech decoding process involved deductive tracking of temporal dynamics over a few spectral regions; thus he concluded that the most important role played by the auditory system was to provide segmental information based on the perceived changes of temporal dynamics, for example transitions. In general, it appeared again to be essential to keep reliable information despite syllabic segmentation to understand spoken language.

Also [Pfitzinger, 1996] and [Pfitzinger, 1998] suggested to refer to the syllable as main unit of speech perception. The researcher demonstrated that the perceived speaking rate was more closely related to the local speaking rate measured in syllables per second than to the one measured in phones per second. However, he proposed to still take into account speaking rate measures based on phones per second since according to his observations a combination of both syllable and phone rate reflected the perceived speaking rate best. In this regard, the author found that syllable rate had a part of 54% in perceived speaking rate whereas phone rate had a part of 46%. Already before, [Portele et al., 1994] came to a similar conclusion when evaluating possible scenarios for duration manipulation in speech synthesis: By means of perceptual evaluation they detected that syllable based duration manipulation revealed better results when compared to natural speech than phone based duration prediction did. Also with regard to the inclusion of coarticulatory phenomena the syllable seemed to be the more probable unit of speaking rate perception since such phenomena usually occur within a syllable frame and much more seldom across syllable boundaries (cf. [Greenberg, 1996]). This view was also supported by other research, for example [Keller and Zellner, 1996], [Widera and Portele, 1999], [Widera, 2003].

In a follow up study on local speech rate perception, [Pfitzinger, 1999] revealed that the previously suggested estimation of the local speech rate based on a linear combination of local syllable rate and local phone rate ([Pfitzinger, 1996], [Pfitzinger, 1998]) probably was not that well-suited to describe speaking rate perception as previously assumed. Since [Kohler, 1986] reported an influence of the fundamental frequency level and its movement on speaking rate perception, [Pfitzinger, 1999] included them in a new linear combination model. The new linear combination was well correlated with the perceptual local speech rate. However, introducing F0 measurements did not increase the accuracy of the model. Additionally, the duration of speech stimuli was found to have a strong influence on the perception of speech rate as well (cf. [Pickett and Pollack, 1963]). Segments of less than 500 ms duration made the perception of speaking rate more difficult whereas segments of more than 700 ms duration contained too much variation of speaking rate to provide a consistent rate perception. Moreover, [Pfitzinger, 1999] observed that the shorter the stimulus the faster the perceived speech rate; he called this the "perceptual overshoot phenomenon". Again, this did not hold for stimuli longer than 625 ms. All subjects had a consistent notion on how to assess speech rate; however, also here the speech rate judgment was different when based on syllable rate as opposed to phone rate.

[Den Os, 1985] investigated the perception of speech rate of Dutch and Italian utterances. In her research, the author evaluated several correlations between linguistic syllables per second, based on orthography, phonetic syllables per second, based on the number of actually produced syllables, and phonetic segments per second, referring to actually produced segments, as well as perceived speaking rate evaluated by means of rate judgments of normal, monotone, and unintelligible utterances. Moreover, short term judgments as opposed to global tempo judgments were investigated. An additional topic was the influence of the intonation on the perceived tempo as well as the question whether human listeners were still able to judge speaking rate when spectral information was missing, based on prosodic factors like intonation, intensity, and rhythm only. Despite the manifold analyses conducted, [Den Os, 1985] found no clear preference for nor an advantage of a specific speech rate measure: The coefficients for unintelligible utterances were lower than for normal utterances, but still significant. The best fitting measure for unintelligible Italian was phonetic syllables per second, whereas for unintelligible Dutch, Italian listeners showed no preference. For Dutch listeners, however, coefficients dropped for unintelligible Dutch. In contrast to the author's expectation, Dutch sounded faster than Italian when judged by Dutch listeners. It was concluded that judgment probably was based on syllable structure or even smaller units, and thus was influenced by the listener's familiarity with the respective syllable structure. All in all, phonetic syllables per second fit least for tempo judgments. A pairwise comparison of different kinds of judgments showed that a lack of intonational information had no effect in the listener's own language. However, a lack of spectral information hindered tempo judgment in both languages. Acoustic cues thus were used in the same manner by both speaker groups at normal speech rate. An additional scaling experiment revealed that judging speaking rate on a seven point scale was not possible for unintelligible utterances. Furthermore, monotone utterances were generally perceived as faster, presumably due to less structure in the signal. The latter observation might be an indication in the direction of the preferences of blind and visually impaired people using fast speech output on a daily basis as reported by [Fellbaum, 1996]. The author stated that generally speaking monotonous utterances were preferred over more lively intonation by this user group. This aspect is more closely investigated in the questionnaire displayed in chapter 5.

4.2 Artificial Fast Speech Perception

The following section takes a closer look at the perception of artificial fast speech, distinguishing between time-compressed natural speech (section 4.2.2) and synthesized (fast) speech (section 4.2.3). At first, an overview is given over methods of perceptual evaluation of (synthetic) speech in section 4.2.1. An evaluation of natural fast speech will take place during speaker selection (cf. section 6.2). The perceptual evaluation of time-compressed natural (fast) speech will then play an important role with regard to the evaluation of corpus recordings conducted (cf. section 7.1). The learnability of listening to fast and/or synthetic speech as well as differences between different listener groups in the perception and judgment of synthesized fast speech will be discussed later on in chapter 4.3. When it comes to the analysis of fast speech synthesized for the current project, experimental results will be analyzed with regard to the abilities of subjects and their experience with listening to fast speech (cf. chapter 8.2).

4.2.1 Perceptual Evaluation of Synthetic Speech

When perceptually evaluating synthetic speech several aspects need to be considered. A perceptual evaluation can be seen as a complementary part of an acoustic evaluation of speech production data (cf. remarks related to duration prediction and its evaluation in chapter 3.2.1). A distinction often made is the one between objective (acoustic) as opposed to subjective (perceptual) measures. An evaluation can be done on a global level by interviewing listeners, or on a diagnostic level by developers. Especially in the case of speech applications mostly developers systematically search for system errors [Hess et al., 1997], [van Santen, 1993], [Pols and Jekosch, 1994], [Lemetty, 1999], [Möller, 2000]. A perceptual evaluation has several dimensions as well: The most important one is *intelligibility* [Hess, 1992], [Hess et al., 1997]. Intelligibility is treated as a quantitative measure: The amount of speech units understood correctly is the evaluation criterion. Intelligibility can be measured by so called "recall" or "close shadowing" tasks, or even on a more global level by "characterization" through listeners [Altmann and Young, 1993], [Dupoux and Green, 1996], [Arons, 1992]. Also the analysis of the word error rate ("WER") gives a good picture of speech intelligibility [Arons, 1992]. More fine-grained intelligibility tests are the Cluster Identification test ("CLID", [Jekosch, 1992]), the Diagnostic Rhyme test ("DRT", [Voiers, 1977]), the generation and evaluation of Semantically Unpredictable Sentences ("SUS", [Benoit et al., 1989]), and the Modified Rhyme test ("MRT", [Sotscheck, 1982]). More recently, also a new method proposed by the Acoustical Society of America's TTS Technology Standards committee (S3-WG91; cf. [Acoustical Society of America, 2013]) is applied [Stent et al., 2011]. A detailed overview over intelligibility evaluation methods and their applicability can be found in [Jekosch, 2005]. It is important to note that intelligibility measures do not necessarily reflect quality judgments. To evaluate the intelligibility of the fast and ultra-fast speech generated for the current project and evaluated as described in section 8.2, Semantically Unpredictable Sentences ("SUS") are the method of choice. Further explanations on methods applied and evaluation results can be found in the referred section.

In addition to intelligibility, the *comprehensibility* as well as the *naturalness* of synthetic speech are important evaluation criteria [Hess et al., 1997], [Fellbaum and Höpfner,]. When evaluating comprehensibility, the measures are less exact than for intelligibility: Subjects usually are asked to summarize and/or tell in their own words what they heard [Arons, 1992], [Hess et al., 1997]. When evaluating naturalness, the results become even more subjective since it is already hard to define what exactly naturalness is. Most naturalness ratings are based on a one-dimensional scale. Based on the recommendations published by the *International Telecommunication Union* (henceforth "ITU") in 1985 ([International Telecommunication Union, 1994], cf. below), this one-dimensional scale is designed to reflect the Mean Opinion Score ("MOS") about certain aspects of the generated speech [van Heuven and van Bezooijen, 1995], [Krause and Braida, 2002]. Next to scalable judgments, also relative preferences are collected in so called "preference tests" [Hecker and Williams, 1966], [Wolters et al., 2010], [Loizou, 2011]. Nonetheless, according to [Hawkins et al., 2000] a one-dimensional scale doesn't make too much sense to evaluate naturalness. The authors suggest to define the degree of naturalness as a function of intelligibility where the category rated best should be describable by an utterance like "as easy to understand as natural (human) speech". This recommendation already shows that specific characteristics are not easily differentiated or judged independently from each other; often, a high correlation between different assessment criteria is observable, especially if judgments are collected from naive listeners [Pols and Jekosch, 1994], [Syrdal et al., 1998], [Hawkins et al., 2000], [Alvarez and Huckvale, 2002], [Jekosch, 2005], [Tucker and Whittaker, 2006]. The difficulty different listener groups show in separating between these categories will also be discussed later on in section 4.3, and investigated in the perceptual evaluation conducted for the current research presented in chapter 8.2.1.

Additional dimensions like overall quality, prosody, stress patterns, and intonation are of interest as well when speech quality is perceptually evaluated [Sonntag, 1999]. When conducting a perceptual evaluation of the quality of the output of five German speech synthesis systems, [Kraft and Portele, 1995] noted only two dimensions of perceptive space quality represented by prosodic and segmental attributes. Previously, [Pols, 1989] defined four broad speech quality assessment categories. First global techniques, addressing acceptability, preference, naturalness, and usefulness of the system. Second, diagnostic techniques to evaluate the quality of segmental units, intelligibility, and prosody. A third category proposed is objective techniques: Measurement of metrics like Speech Transmission Index ("STI") and Articulation Index ("AI"). The forth category would then be application specific techniques: How does the system/speech perform in applied domains? Earlier, [Pisoni, 1981] had even suggested to evaluate (synthetic) speech perception in ten areas to cover for all its possible aspects. To allow for a more systematic comparison, the ITU published a detailed specification of a method for subjective performance assessment of the quality of speech of voice output devices in 1985 [International Telecommunication Union, 1994]. It was recommended that subjects expressed their opinion on certain speech characteristics on one or more rating scales reflecting Mean Opinion Scores (also called "CE" (category estimation)) after having answered specific questions on the information contained in the presented utterance instead of using preference tests. The results could then serve as measures of the perceived quality in several aspects. According to the ITU, this method also takes into account two aspects influencing speech rating fundamentally: the performance of the system and the attitude of the listener. Thus, the method was recommended since it could also cover for judgments of overall system performance as well as of the applicability to a specific task. The overall goal of the ITU-TP85 recommendation was to obtain comprehensive results as well as to improve the description of listeners' perception by using multiple scales on different aspects. The utterances used for evaluation were to be related to practical applications. Subjects thus were urged to pay attention to the information contained in the presented stimuli before expressing their opinion, mostly on a five-point Mean Opinion scale. Tests with stimuli on sentence level, even with SUS, were seen as especially useful to evaluate the intelligibility of a system.

[Sityaev et al., 2006] compared the ITU-TP85 standard to other methods used to evaluate TTS Systems and noted that it had neither been widely accepted nor largely used to its full extent since being published. Nevertheless, the authors agreed that the two major aspects of TTS system evaluation were intelligibility, usually evaluated by means of SUS, and naturalness, mostly judged by means of MOS. Thus, the goal of their research was to investigate whether the ITU test as a whole could provide a better performance measure than SUS and MOS alone if applied to several aspects of speech. The researchers found that subjects preferred to have six scales to judge certain aspects of speech instead of the nine scales suggested by the ITU. Therefore, they decided to henceforth divide their evaluation into two parts: One to evaluate intelligibility aspects, and one to investigate other speech quality aspects to make judgments easier and more independent from each other. Moreover, instead of using longer passages from one genre as recommended by ITU-TP85, they applied the SUS method to collect more rigorous and informative data about misrecognitions. Furthermore, the systematic differences in grammar of the SUS [Benoit and Grice, 1996] were expected to reveal problems in different areas like syntax or prosody. In contrast to the findings of [Alvarez and Huckvale, 2002], [Sityaev et al., 2006] noted that not all scales were correlated. However, a striking observation was that the system which scored highest for intelligibility scored lowest for naturalness and overall quality which was in line with [Jekosch, 2005]'s conclusions.

Following the ITU-TP85 guidelines widely, also [Chalamandaris et al., 2010] studied the usability of a specific TTS system by means of MOS regarding naturalness, ease of listening, and clearness of articulation at sentence level. Intelligibility was evaluated by DRT tests on word level. Speech flow and overall listening experience, however, were judged on paragraph level by listeners indicating MOS. Main usability aspects such as effectiveness, efficiency, and satisfaction were evaluated as well by interviewing potential target users. Additionally, the behavior of participants while conducting the evaluation was observed. This way, the researchers revealed that systems which were tailored to specific domains and application requirements achieved higher quality results and better performance. Crucial factors for satisfaction or frustration of the end-users were responsiveness, flexibility and intelligibility of the system. [Chalamandaris et al., 2010] concluded that TTS technology needed to be adapted and customized for dedicated services and tools. The aspects pointed out here as being important for potential users were already revealed by [Scholtz, 2004]: Efficiency, learnability, memorability, error rate, user satisfaction, as well as voice quality and suitability.

[Hammerstingl and Breuer, 2003] already came to a similar conclusion as [Chalamandaris et al., 2010] when evaluating the BOSS system developed at the University of Bonn, the system also used to generate fast speech in the framework of the current project (cf. section 3.1.2). They noted that modern TTS architectures required additional evaluation processes since the cost functions applied often were like a "black box" such that investigators could not foresee the exact unit selection for each and every context. For their evaluation of the performance of proper name synthesis, [Hammerstingl and Breuer, 2004] adapted existing test methods. Stimuli were selected in accordance with being representative for the application scenario (cf. [Wagner et al., 1999], [Sonntag, 1999]). Preference tests in terms of A-B-comparisons as well as an evaluation of the phone error rate detection were conducted. The authors revealed as one of their findings that a higher number of concatenation points let to an increased number of intelligibility errors. Thus, to reduce the number of concatenations [Breuer, 2009] suggested later on to introduce another unit size: "phoxsy units". Their applicability to the current approach of fast speech synthesis is outlined and evaluated in chapter 8.1. A domain specific evaluation, however, will not be conducted since screenreader applications are usually used in an unrestricted domain.

[Demenko et al., 2010] evaluated the quality of Polish unit selection speech synthesis, using the BOSS system as well. Two kinds of perception tests were carried out: Preference tests to investigate synthesized speech obtained by using different versions of speech segmentation, and a MOS test to evaluate the quality of the Polish speech generated. The authors pointed out that after [Möbius, 2000], "an [..] important task is to collect a sufficient amount and type of speech data representative for the target language". It can be assumed that this also applies to specific speaking styles, like the one investigated in the current research. Since no comprehensive intonation model for Polish was implemented yet, the speech synthesized by [Demenko et al., 2010] suffered from issues in prosody. Thus, to enhance the naturalness of the generated output, the authors suggested to additionally develop an intonation model for Polish while focusing on acoustic correlates of stress and accents as well as on the degree and actual position of stress and accents. In contrast to the usual SUS tests, the utterances they used in their MOS test were rather simple: Twenty five "common vocabulary sentences" derived from the topics "Common, Conversation and Natural". Such simple sentences were also used to define the adequate unit size as described in section 8.1.

In 2010, [Möller et al., 2010] conducted an evaluation of different approaches for instrumentally predicting the quality of TTS systems. Their motivation was the fact that the evaluation of synthesized speech was still a frequent and important task but all well-established tests were relying on listeners, and therefore were time-consuming and expensive. Thus, the authors were aiming at the development of instrumental, objective estimates of speech quality. Prediction performance and robustness were to be enhanced by combining HMM-based feature comparison and parametric approaches based on a log-likelihood determination; features extracted from synthesized speech were compared with features derived from natural speech. Parameters were extracted from the signal which captured quality-relevant degradations of the synthesized speech. The results of the evaluation showed that this way auditory quality judgments often were predictable with a sufficiently high accuracy and reliability. In five out of six test cases, correlations higher than 0.8 could be obtained, and further increases could be achieved on a per-synthesizer basis. The latter finding led to the conclusion, however, that the tested approach was more applicable to distinguish between synthesizers instead of between utterances from one synthesizer.

Setting forth this research, [Hinterleitner et al., 2011] investigated the perceptual quality dimensions of state-of-the-art TTS systems. They conducted several pretests to determine suitable attribute scales. The first pretest was designed to collect a broad basis of attributes reflecting auditory features by means of subjects writing down nouns, adjectives, and antonym pairs describing their auditory impression. The result was a list of 28 preferred attributes. The purpose of the second pretest was to narrow down the set of attributes resulting from pretest one. Thus, subjects were asked to use only those attributes which were most relevant for their auditory impression, resulting in 16 remaining attributes. In the concluding main test subjects indicated their overall impression on a continuous scale representing MOS, and single attributes were judged afterwards via a slide. The following multidimensional factor analysis revealed three main factors accounting for 61.47% of the total variance in judgments: naturalness, disturbances, and temporal distortions. As found earlier (cf. [Jekosch, 2005]), all factors were correlated, especially factor one (naturalness) and factor three (temporal distortions). Mapping single factors onto the perceived overall quality revealed that naturalness contributed the most to the perceived quality of the presented TTS signals. However, [Hinterleitner et al., 2011] concluded that modern TTS systems suffered from diverse quality constraints, not only related to insufficient naturalness.

Irrespective of the huge number of possible evaluation categories and characteristics, the perceptual experiments conducted during the current project are restricted to two categories only. Naturalness and intelligibility play the most important role during the investigation of the selected speaker's fast speech (cf. chapter 6.2.2), the evaluation of corpus recordings (cf. section 7.1.1), the determination of the adequate unit size (cf. section 8.1) as well as the evaluation of fast and clear speech SUS synthesized with different synthesis approaches at different speaking rates (cf. chapter 8.2). Moreover, MOS were collected for all stimuli.

4.2.2 Time-compressed Speech Perception

Compared to natural fast speech, the perception of time-compressed speech is more difficult [Foulke, 1971], [Elsendoorn, 1985], [Winters and Pisoni, 2004], [Papadopoulos et al., 2010]. As outlined in chapter 2, in natural fast speech production a severe deterioration of several acoustic characteristics can be observed. Nonetheless, those characteristics are necessary for the correct identification of what has been articulated. Thus, their deterioration causes severe problems in the perception of natural fast speech. Time-compressing natural fast speech will presumably intensify these problems. With regard to the research presented here, it is of interest to compare natural fast speech perception to the perception of time-compressed normal and fast rate speech as well as synthesized (fast) speech. The following section will therefore take a closer look at the perception of time-compressed speech prior to the discussion of the perception of synthesized (fast) speech in chapter 4.2.3.

Already in 1969, [Foulke and Sticht, 1969] noted that connected discourse comprehension decreased slowly up to a word rate of 275 words per minute. Beyond that point, comprehension dropped more rapidly than before. The authors ascribed this phenomenon to a processing overload of the short term memory for fast speaking rates. However, their experiments revealed that repeated exposure to fast speech improved word intelligibility. Also [Heiman et al., 1986] observed that with more than 50% of compression important, non-redundant information was lost after word intelligibility decremented significantly. Following [Arons, 1992], [Foulke, 1971] found that intelligibility of words compressed to 10% of their original duration was still given whereas comprehension of connected texts decreased at 50% of their original duration already, in accordance with [Heiman et al., 1986]. Also here, repeated exposure to time-compressed speech increased both intelligibility and comprehension. The authors concluded that intelligibility was more resistant to degradation as a function of time-compression than comprehension. However, if immediate memory span was exceeded, intelligibility decreased as well. The following conclusion was also derived from the detection that for listeners who were allowed to stop time-compressed recordings at any time, the interval from start to stop was almost proportional to the increase in speaking rate.

> While time and/or capacity must clearly exist as limiting factors to a theoretical maximum segment size which could be held [in short-term memory] for analysis, speech content as defined by syntactic structure is a better predictor of subjects' segmentation intervals than either elapsed time or simple number of words per segment. This latter finding is robust, with the listeners' relative use of the [syntactic] boundaries remaining virtually unaffected by increasing speech rate. [Wingfield and Nolan, 1980] after [Arons, 1992].

When investigating factors affecting perceptual adjustment to time-compressed speech, [Altmann and Young, 1993] aimed at discovering the mechanisms underlying this adaptation. Another goal of their research was to identify the processing unit which the human recognition system attempted to recover during this adaptation. The investigators found that the observable adjustment was not driven by lexical level word recognition. Instead, they assumed that rather certain sub-lexical units or supra-segmental regularities formed its basis. Moreover, an influence of prior exposure to fast speech on the ability to adapt to fast speaking rates was detected. The authors concluded that adaptation to time-compressed speech was more than just a short-term retuning. According to [Altmann and Young, 1993], prior studies had revealed that the intelligibility of highly compressed speech strongly correlated with the plausibility of the presented material. However, they discovered that in their experiments adaptation also took place for SUS. In general, adaptation to time-compressed speech showed to be language-dependent in terms of dependency on language-specific factors. This finding was confirmed by [Pallier et al., 1998] who evaluated perceptual adjustment to time-compressed speech across different languages. In accordance with [Voor and Miller, 1965] they found that the performance of listening to artificially accelerated speech improved in the course of ten to fifteen sentences. This was observed even for monolingual speakers when listening to strongly related languages like Spanish versus Catalan. [Pallier et al., 1998] pointed out that the listener's processing apparatus had been designed to compensate for input's instability like noise etc. by nature. Therefore, compensation took place almost instantaneously and effortless. However, in suboptimal conditions like noisiness or increased speaking rate, a larger processing time frame was necessary. Thus, the researchers evaluated whether the perceptual adjustments involved the processing systems that map the acoustic information onto the proper lexical representation. In that, they found that speakers mostly relied on adaptation processes specific for their native language, especially on phonological properties. The authors concluded that adaptation to one specific language was of little use in another language, unless the other language revealed similar phonological properties.

Regarding adaptation and habituation to accelerated speech in general, however, one can find differing observations in the literature. Where [Orr et al., 1965] (after [Arons, 1992]) claimed that a substantial increase in intelligibility was only achievable after eight to ten hours of training, [Voor and Miller, 1965] or [Carlson et al., 1976] stated that it took their subjects eight to fifteen sentences rather than several hours to adapt to fast speaking rates. Moreover, [Beasley et al., 1976] noted that listeners felt uncomfortable returning to normal speech (cf. [Dupoux and Green, 1996]) after repeated exposure to accelerated speech. [Dupoux and Green, 1996] put their focus on the effects of talker and rate changes in highly compressed speech. Perceptual adjustment was investigated on a number of different levels during fast speech processing. As noted by other researchers before, the authors found perceptual adaptation to occur over a number of sentences, thus in a rather short time frame. The gradient of adjustment was dependent on the compression rate. Abrupt changes either in talker or in compression rate had only little effects on adaptation. The authors assumed that different instances of events, for example spoken vowels, were normalized during perception (cf. section 4.1.1), also with regard to rate induced variation. Thus, normalization was seen as immediate response to local speech rate variation. An adjustment to talker specific speech including accented or dialectal speech was found to take place after a few minutes as well. According to [Schwab et al., 1985] (after [Dupoux and Green, 1996]), this also applies for synthetic speech. A remarkable finding in the context of the current research is that with increasing exposure subjects reported fast speech to sound less unnatural and better to understand.

[Dupoux and Green, 1996] documented the degree of perceptual adaptation as a function of the amount of experience with compressed speech in more detail. In doing so, they found a significant increase in number of content words correctly recalled across four different sentence sets presented in successive test sessions. This phenomenon was not speaker-specific. However, the authors noted that the adjustment process required more time when the compression rate was higher. In accordance with the Motor Theory of speech perception (cf. section 4.1.1), the improvement was specific to compressed speech and not transferable to other accelerated acoustic signals. The researchers concluded from their observations that the adjustment to compressed speech may have been the result of two processes operating simultaneously: a short-term adjustment to local speech rate parameters, and a longer-term adjustment reflecting a more permanent perceptual learning process. Furthermore, it was hypothesized that adaptation operated on an abstract level such that acoustic differences between talkers did not matter.

In a more recent study, [Adank and Janse, 2009] investigated the processes involved in perceptual learning of time-compressed versus natural fast speech. They found that listeners' performance on natural fast sentences was significantly poorer than on normal rate sentences, but performance on time-compressed sentences was not. Additionally, transfer of learning was observed when time-compressed speech was presented before natural fast speech, but not vice versa. In accordance with [Trouvain, 2004] and others, [Adank and Janse, 2009] noted that speakers increased their speaking rate in a non-linear fashion. Therefore, they concluded that listeners were forced to permanently normalize for variations in speech rate (cf. [Dupoux and Green, 1996], [Pallier et al., 1998], section 4.1.1). However, the authors also pointed out that it was questionable whether time-compressed speech itself provided a useful model for perceptual adaptation to specific characteristics of naturally produced fast speech which seemed to be more difficult to process (cf. [Janse, 2003a], [Janse, 2004]). They based this assumption on their finding that improvement over trials was higher when time-compressed speech was presented after natural fast speech, but better performance on natural fast speech was achieved when those stimuli were presented after time-compressed stimuli. Additionally, the researchers evaluated the "speed of language comprehension" ("SCOLP") on the basis of the percentage of correctly identified words. They noted an increase of the processing speed after exposure to time-compressed speech which they interpreted as a sign for perceptual learning. In general, responses got faster for time-compressed stimuli as well as for natural fast speech stimuli over time, but response times were higher for natural fast speech, mainly in the initial two to three test blocks. [Adank and Janse, 2009] summarized their findings as follows:

> Importantly, whereas adaptation to time-compressed speech did not show up as improved accuracy over trials, it was found in decreased response times over trials. Adjustment to natural fast speech was found both in improved accuracy and somewhat decreased response times over trials. [Adank and Janse, 2009].

Furthermore, [Adank and Janse, 2009] observed an improvement also in the normal rate condition when comparing SCOLP results for the first half versus the second half of the presented stimuli. To conclude with, they pointed out that so called ”practice effects” were a possible, competing explanation to perceptual adaptation theories. In the ”Reverse Hierarchy Theory” by [Ahissar and Hochstein, 2004] (after [Adank and Janse, 2009]), for example, perceptual learning was defined as practice-induced improvements in the ability to perform specific perceptual tasks, not as adaptation of the perceiving system. However, the investigations discussed here clearly showed that in general, an adaptation to time-compressed speech took place, even if it required longer exposure to the speaking style in question than for natural fast speech. Whether and how this adaptation has an influence on the quality judgments gathered in the connection of the research presented here will be detailed in chapters 7.1 and 8.2.

4.2.3 Synthesized Speech Perception

The perception of synthesized (fast) speech is more difficult than the perception of natural (fast) speech as well, and even more complex than the perception of (linearly) time-compressed speech. [Winters and Pisoni, 2004] and [Papadopoulos et al., 2010] pointed out that synthesized speech was less intelligible than natural speech in general. This had already been observed by [Pisoni, 1981]: They found that synthetic speech required more cognitive resources revealed by longer reaction times and more numerous errors in close shadowing than natural speech. Moreover, recall performance was worse (cf. [Luce and Pisoni, 1983], [Bailly, 2003], [Winters and Pisoni, 2004]). [Schwab et al., 1985] found that synthetic speech produced by rule lacked both the rich variability and the acoustic-phonetic cue redundancies of natural speech. Also the lack of appropriate prosodic information was a disadvantage in perception. Furthermore, the authors observed that synthetic speech generated by concatenation may damage the perceptual quality by introducing discontinuities in the speech signal, but have advantages compared to synthetic speech generated by rule because it includes robust and redundant sets of perceptual cues to individual segments in the signal which is important for a better perception. This hypothesis was picked up in the considerations made about the work presented here: The use of unit selection speech synthesis is expected to enhance intelligibility and naturalness of fast synthesized speech. Whether this is really the case will be discussed in chapter 8.

[Winters and Pisoni, 2004] investigated the perception and comprehension of synthetic speech produced by applying durational rules. They evaluated segmental intelligibility, word recall, lexical decision, sentence transcription, and sentence comprehension. Moreover, the authors aimed at accounting for perceptual differences in terms of acoustic-phonetic characteristics. They confirmed their previous findings regarding the lack of variability contained in natural speech in speech synthesized by rule. Furthermore, speech synthesized by rule provided fewer redundant cues, revealed only highly simplified coarticulation, and also showed a lack of appropriate prosody contours. [Winters and Pisoni, 2004] pointed out that redundancy and variability were to be seen as fundamental properties of natural speech, not as a disturbing noise to be normalized during perception. Additionally, they stressed that from their point of view the intelligibility of synthetic speech depended greatly on the type of synthesizer being used to produce it. When comparing formant synthesis to concatenative synthesis, earlier studies showed that formant synthesis was (slightly) more intelligible than concatenative synthesis ([Hustad et al., 1998] after [Winters and Pisoni, 2004]). Nevertheless, more recent studies like the one conducted by [Venkatagiri, 2003] (after [Winters and Pisoni, 2004]) where the intelligibility of four different synthesis systems compared to natural speech was evaluated in multi-talker bubble noise revealed that natural speech utterances were much easier identified than synthesized speech. Moreover, concatenative synthesis was significantly more intelligible than formant-based synthesis. Although formant-based synthesis outperformed concatenative synthesis on formant intelligibility, significantly more listeners made errors on consonant identification. [Winters and Pisoni, 2004] hypothesized that this was the case because concatenative synthesis kept the important consonantal transitions occurring in natural speech (cf. section 2.2.2) whereas in formant synthesis those were produced in an artificial manner. However, focusing on the perception of concatenative synthesis only, [Fowler, 2005] found that listeners were more disrupted when spliced and re-generated syllables provided misleading acoustic information about the forthcoming vowel than when the information was accurate. She concluded that destroyed transitions and coarticulatory effects made the perception of speech generated by unit selection TTS more difficult. This question is picked up in the discussions about advantages and disadvantages of the different synthesis systems applied for the current research in chapter 8.

Returning to the question whether the perception of synthetic speech required more cognitive resources, probably ascribable to poor segmental intelligibility, [Winters and Pisoni, 2004] conducted a speeded lexical decision task which showed that the lexical decision was always slower for synthetic speech stimuli, no matter whether the stimulus was an existing word or not. The authors assumed that this was the result of a greater familiarity with natural speech. Referring to [Dupoux and Green, 1996], [Winters and Pisoni, 2004] pointed out that listener response times indeed decreased over a five-day training period for strings produced by synthetic speech but that response times never reached the same level as the ones for natural speech. Moreover, listeners recalled fewer words from synthetic word lists. However, the perception of synthetic words presented in meaningful context has been found to be significantly better than the perception of synthetic words presented in isolation, presumably reflecting the influence of higher level linguistic information. The researchers noted that here the poor-quality synthetic speech may have misled listeners and thereby been less informative than no signal at all. Additionally, the perception of words in sentences worsened if the sentences lacked semantic coherence and predictability, again even more for synthetic than for natural speech. [Winters and Pisoni, 2004] referred to the numbers found by [Pisoni and Hunnicut, 1980] who revealed 97.3% correct word identification for natural speech but only 78.7% for synthetic speech because the listeners developed misleading expectations from the presented higher-level semantic information. Nevertheless, comprehension of synthetic speech was shown not to be worse than that of natural speech. [Winters and Pisoni, 2004] summarize their explanations with the conclusion that ”it may take longer to process synthetic speech than natural speech, but the final levels of comprehension achieved for both types of speech are ultimately equivalent.”

Despite, the improvement in perceiving fast and/or synthetic speech may not continue indefinitely: For example [Venkatagiri, 1994] observed a so called ”ceiling effect”. Elsewhere, this phenomenon is referred to as reaching a ”plateau” in perception improvement [Adank and Janse, 2009]. Based on their finding that improvement on a natural voice was greater than improvement on a synthetic voice, [Slowiaczek and Pisoni, 1982] concluded that the ability to comprehend synthetic speech would always be worse than comprehension of natural speech, in contrast to [Winters and Pisoni, 2004]. Furthermore, they hypothesized that improvement was due to exposure as such. The observed improvement in comprehension ability was seen as domain specific: Following [Slowiaczek and Pisoni, 1982], the performance improved due to the development of an increased proficiency in interpreting lower-level acoustic-phonetic details of synthetic speech. In accordance with [Carlson et al., 1976], [Slowiaczek and Pisoni, 1982] finally noted that the observable training effects had long term benefits also for synthetic speech comprehension.

When conducting a multidimensional analysis of factors influencing synthetic speech quality ratings, [Hinterleitner et al., 2011] found three underlying quality dimensions influencing listeners' judgments the most: naturalness, disturbances, and temporal distortions of prosodic characteristics. That the intelligibility of synthetic speech was to some amount dependent on the appropriateness of prosodic cues had already been observed by [Slowiaczek and Nusbaum, 1985]. Also [Sanderman and Collier, 1997] and [Winters and Pisoni, 2004] noted that synthetic sentences with appropriate prosodic contours facilitated comprehension whereas inappropriate contours made comprehension more difficult. [Carlson et al., 1979] observed as well that duration and fundamental frequency contours severely deviating from natural speech decreased intelligibility significantly. However, the researchers also noted that not all subjects considered the natural speech reference as the best version (cf. [Portele, 1997], [Black and Tokuda, 2005]). Furthermore, [Carlson et al., 1979] noticed a general correlation between naturalness ratings and measurable differences in duration, but pointed out that nonetheless physical distance was no reliable predictor of perceptual distance. The authors' conclusion was that rules modifying the duration of a segment as a function of syntax and segmental context were of significant importance for both naturalness and intelligibility of synthetic speech. Moreover, their results clearly indicated that correct segmental durations resulted in significantly better intelligibility and naturalness which was one of the reasons to conduct the evaluation presented in section 7.2.2). Additionally, they revealed that basically a correlation between intelligibility and naturalness existed.

[Huggins, 1979] evaluated the effects of inappropriate temporal relations within speech units on the intelligibility of synthetic speech as well. She found that badly disturbed speech timing, reflected in speech either being too slow or containing inappropriate pauses, led to an incorrect segmentation of the respective utterance causing a severe loss of intelligibility. Also the pattern of stressed syllables seemed to be important for correct perception. When stress was wrongly assigned in synthesized sentences, the intelligibility of words fell from 85% to 50%, and the percentage of comprehended sentences decreased from 75% to 25%. However, in case a listener knew the content of an utterance s/he was not able to estimate the effect of a particular timing distortion on speech intelligibility. At the end of her explanations, [Huggins, 1979] pointed out that the human perceptual apparatus is to be seen as very good in filling in missing information, but that it is very bad at discarding extraneous information. Therefore, the stress pattern of a word or phrase was of critical importance to its correct recognition.

In 2002, [Janse, 2002] conducted experiments on the perception of time-compressed naturally produced and artificially generated words. She hypothesized that the hyper-articulation found in the synthetic speech she employed was attributable to the fact that the diphones used were derived from stressed syllables. The author assumed that this circumstance would enhance the intelligibility of synthetic speech at fast rates because of inherent segmental redundancy and despite of high unnaturalness. This hypothesis showed not to be true. Moreover, the advantage of natural speech did not decrease after time-compression. Nevertheless, detection times tended to become shorter in general, probably due to shorter syllable and word duration in fast speech, although synthetic diphone speech was rather blurred at fast speaking rates. The higher intelligibility of fast natural speech was attributed to segmental intelligibility, lexical redundancy, and context information. In accordance with previous observations, [Janse, 2002] additionally noted differences in processing speed between natural and synthetic speech when conducting a phoneme detection task. Longer response times for synthetic speech were observed. A higher processing load for fast synthetic speech rates was assumed to be the reason for this. Again, non-assimilated and unreduced forms were recognized easier in fast speech than assimilated forms. The sparsity of phonetic cues in synthetic speech was seen as an additional disadvantage. Furthermore, misleading coarticulatory acoustic cues and the absence of variation in speaking effort in synthetic speech showed to be even more severe in time-compressed synthetic speech (cf. [Winters and Pisoni, 2004]).

In a follow-up study, [Janse, 2003b] pointed out again that the entire temporal structure of fast speech differed significantly from the one of normal speech. By means of perceptional experiments she revealed that the intelligibility of fast spoken words with a temporal pattern similar to natural fast speech was lower than fast words generated by linearly compressing normal rate speech. The author concluded that linearly compressed speech comprised some perceptional advantages on temporal and segmental level compared to natural fast speech. Moreover, [Janse, 2003b] found that the segmental as well as the temporal changes observable in fast speech were attributable to articulatory restrictions and did not have any communicative function. The less a stimulus deviated from its canonical form, the better it was perceived. Additionally, it was also shown that if natural speech was compressed up to 65% of its original duration it was still "perfectly intelligible" [Janse, 2003b]. Obviously, the inherent natural acoustic transitions kept the speech intelligible even at fast tempo but the content needed to be semantically or pragmatically predictable to be understood. And even if the temporal compression was further intensified and the compressed utterances comprised only 35% of their original duration, they remained comprehensible in the majority of cases (53%, cf. [Janse et al., 2003]). A complementary observation was later made by [Lebeter and Saunders, 2010] who stated that linearly time compressed speech was generally more intelligible than linearly time compressed synthesized speech.

The intelligibility and naturalness of fast speech generated by different synthesis systems will be evaluated in chapter 8.2. However, the experiments discussed there will not include natural fast speech. The intelligibility of natural fast speech produced with different speaking styles will be discussed in chapter 6.2.2. Afterwards, the intelligibility and naturalness of linearly time-compressed speech are examined in chapter 8.2.

4.3 Listener Dependent Aspects of Speech Perception

As pointed out in section 4.2.1, the perception of speech - and in particular the perception of speaking rate and fast speech - is influenced by individual listener characteristics. Aside from mother tongue and geographic origin, listener judgments were found to be influenced by several other factors like listeners' age, expectation, familiarity with the presented material, language proficiency, accuracy of hearing, motivation, and even emotional condition [Möller, 2000], [Jekosch, 2005], [Black and Tokuda, 2005], [Syrdal et al., 2012]. To overcome the issue of listener dependent judgments, [Black and Tokuda, 2005] chose three different listener groups to participate in their evaluation of different speech synthesis systems. The authors expected the different listener groups to have different goals: It was assumed that ”speech experts” (that is researchers interested in speech synthesis) would vote most careful, ”volunteers” recruited through the internet more randomly, and ”undergraduate students” probably were less motivated and thus less reliable despite payment. Independent from the respective listener group, [Black and Tokuda, 2005] observed that ”better” was not the same for everyone all of the time (cf. [Portele, 1997], [Brinckmann and Trouvain, 2003]). However, [Bennett and Black, 2006] noted that speech expert listeners were generally better at understanding synthetic speech which became apparent in a lower word error rate. Furthermore, this listener group also liked synthetic speech more than other groups which the authors attributed to some habituation effects. The latter is also of interest with regard to the current research, in particular the familiarity with synthetic speech as such, as well as regarding familiarity with a certain speaker or source of speech which can have a significant influence on individual ratings (cf. chapter 8.2.1). [Tucker and Whittaker, 2006], for example, also noted that known voices were easier to understand than unknown ones (cf. [Pisoni, 1993], [Bradlow et al., 1995], [Traunmüller, 2000]). Similar familiarization effects were observed by [Pisoni, 1981], [Luce and Pisoni, 1983], and [Sonntag, 1999]. And it was for the same reason that [Bond and Feldstein, 1982] decided to scale their evaluation categories of speaking rate dependent on the experience of listeners which they described as ”trained” or ”untrained”. This distinction between listener groups was taken over for the investigation of synthesized fast speech described in section 8.2. During his research in non-uniform time-scaled speech perception, [Höpfner, 2007], [Höpfner, 2008] asked sighted, visually impaired, and sightless subjects about their perception of different speaking rates. Along with the visual ability of the respective subject the author observed a dependency of intelligibility ratings on the duration of the training phase as well as on sentence order.

When proposing a method to evaluate TTS output at higher levels of linguistic organization instead of segmental level, and after testing the comprehensibility of synthesized speech at paragraph level, [Jongenburger and van Bezooijen, 1992] stated that ”[i]n any case, the results of the present study suggest that comprehensibility of natural and synthesized texts does not have to be tested separately for sighted and non-sighted people, results for the one group being generalizable to the other.”. This observation will be investigated further as well in relation with the perceptual evaluation of fast synthesized speech elaborated in chapter 8.2. In contrast, [Trouvain, 2006] and [Papadopoulos et al., 2010] noted that the blind and visually impaired were better at understanding synthesized speech than sighted individuals. In a follow-up study of differences between sighted and visually impaired listeners on the comprehension of synthetic versus natural speech, however, [Papadopoulos and Koustriava, 2015] found that both individuals with and without visual impairments performed at a similar level in the comprehension of texts that were presented via synthetic and natural speech. The findings indicated that local difficulties related to intelligibility did not affect overall comprehension. It seemed that context cues provided through the content helped participants in identifying and comprehending utterances more effectively. Moreover, the results revealed no significant differences between sighted participants and participants with visual impairments regarding the comprehension of natural and synthetic speech as well.

[Jongenburger and van Bezooijen, 1992] evaluated the acceptability of different aspects of synthetic speech as a function of experience. They were especially interested to answer the question whether experience with a certain system enhanced or inhibited the evaluation skills of the listener and whether a carry-over effect to other kinds of (synthetic) speech existed. Acceptability was evaluated in terms of ten different criteria. Results were found to be rather redundant in showing similar patterns of significant effects. Again, intelligibility and naturalness were shown to be the best fit to two groups of criteria. The observation that exposure to high-quality output did not raise perceived intelligibility whereas exposure to lower-quality output did led the authors to the conclusion that listeners indeed were able to learn interpret segmental characteristics of a particular system. However, repeated exposure did not reveal a more positive perception nor rating, in contrast to intelligibility judgments. This aspect of fast (synthesized) speech perception will be discussed in chapter 8.2 as well.

In their studies, [Schwab et al., 1985] posed the question to which extent findings about increased perceptual performance based on repeated exposure to a certain kind of speech could be generalized. They presented their listeners with a huge diversity of training stimuli including a notable amount of acoustic-phonetic variability which increased performance significantly. Nonetheless, also they found that there were limitations of improvement , as already discussed in section 4.2.3. Comparing different listener groups, the authors observed that the performance of ”expert listeners” who had extensive practice listening to the specific form of synthetic speech that their computers produced was far more advanced than for other listeners. Like [Adank and Janse, 2009], [Schwab et al., 1985] also noted that only the training on synthetic speech improved listeners' performance. Moreover, the researchers revealed that appropriate prosody did not have a consistent effect on individual word intelligibility, but rather on naturalness ratings of complete discourse. They concluded that prosodic information in synthetic speech was useful when the syntactic structure of a sentence was not predictable. However, the authors pointed out that synthetic speech was to be seen as impoverished speech which was of limited utility in noisy environments. In accordance with [Slowiaczek and Pisoni, 1982], they hypothesized that "expert listeners" may have developed better abilities to extract acoustic-phonetic information from synthetic speech signals. Nevertheless, the researchers proposed to reduce the number of synthesized messages for cognitively demanding tasks.

In a whole series of investigations, [Trouvain, 2006], [Trouvain, 2007], [Moos and Trouvain, 2007], [Moos and Trouvain, 2008] and [Moos et al., 2008] revealed that blind subjects were able to understand synthetic speech at speaking rates way beyond natural speaking rate and not intelligible for (untrained) sighted people anymore. Thus, the tempo of speech intelligible to sighted listeners was much slower (up to 14 syllables per second) than for blind listeners (up to 22 syllables per second). [Moos and Trouvain, 2007] categorized such "super human speech rates" as "ultra-fast" (cf. chapter 2.1). They summarize their findings as follows:

> That means that the non-blind were still able to follow the message at a tempo which corresponds to the most extreme rates of human speech production, whereas the blind subjects were able to go well beyond this point. Interestingly, this result holds true for synthetic speech generated with a formant synthesizer. [...] Fast speech at rates higher than 10 s/s produced with diphone synthesis was nearly as unintelligible for the blind as for the sighted persons. [Moos and Trouvain, 2007].

Additionally, the researchers noted that the ability to comprehend ultra-fast synthetic speech was not transferred to the comprehension of compressed ultra-fast natural speech (cf. [Schwab et al., 1985], [Adank and Janse, 2009]). Experimental results revealed that the observable training effect reached a plateau after ten minutes for compressed natural speech (cf. [Voor and Miller, 1965], [Adank and Janse, 2009]), or after five days for synthesized sentences (cf. also [Reynolds et al., 2002]). However, [Trouvain, 2006], [Trouvain, 2007] noticed that listeners got exhausted after approximately 30 minutes of listening to extremely fast synthetic speech. That synthetic speech generated by means of formant synthesis at speaking rates of up to 17.5 syllables per second was still comprehensible to their blind subjects was ascribed to the intense and long-term training these subjects had undergone. For diphone synthesis, in contrast, comprehension declined for both listener groups for speaking rates faster than 7.5 syllables per second, but nevertheless there was a difference in comprehension scores between listener groups. From this, [Trouvain, 2007] concluded that diphone synthesis was not more appropriate to generate fast speech than formant synthesis. He assumed that this was related to the many concatenation points occurring in diphone synthesis and suggested to replicate his experiments deploying a non-uniform unit selection speech synthesis system. Moreover, the author also mentioned that the most extreme reading rates of human beings was at approximately 10.5 syllables per second and hypothesized that the rapid decline of comprehension of speech faster than this was no coincidence. These implications will be further discussed in relation to the findings of the findings of the current research examined in section 8.2.

[Dietrich et al., 2013] also investigated whether and how subjects with normal or residual vision could improve their understanding of accelerated speech. Therefore, they asked their subjects to undergo a training period of approximately six months while speeding up the syllable rate of the applied speech synthesis system more and more. Results of the concluding perceptual evaluation revealed that different areas of the subjects' brains were reorganized, and were redistributed differently for different listener groups. Next to this, also individual, distinct strategies of ultra-fast speech processing were observed. The authors ascribed this to the phenomenon of "neuroplasticity". Testing [Trouvain, 2007]'s assumption, the researchers also revealed a difference between listener groups in understanding speech produced at a naturally achievable speaking rate of 8 syllables per second compared to ultra-fast speech generated at 22 syllables per second. Also here, improvement was not homogeneous: Two out of six subjects showed no or only slight improvements after intermediate training on synthetic speech generated at 18 syllables per second.

Similar analyses were performed by [Asakawa et al., 2002] and [Asakawa et al., 2003]. The researchers conducted several perceptual evaluations of the maximum and the most comfortable listening speed for Japanese TTS output for the blind. The subjects were asked to indicate the highest and the most suitable listening rate where the highest listening rate required recognition of at least 50% of the content, and the most suitable listening rate called for recognition of at least 90% of the content with comfortable listening effort. Results showed that advanced blind listeners were able to understand a document read out 2.6-2.8 times faster than the default rate of the TTS. The highest rate often changed depending on the difficulty of the content of the presented material as well as of single sentences and words (cf. also [Höpfner, 2008]). Moreover, [Asakawa et al., 2002] found that the most suitable rate even for novice users was 1.6 times faster than the default TTS rate. The authors concluded that it would be desirable to make the reading rate adjustable for advanced users to improve TTS usability. To not be affected by quality differences among TTS systems and because the applied Japanese TTS did not support rates higher than 900 morae per second, a pre-recorded human voice was used to generate stimuli at highest rates (1300 morae per second). This may have falsified the results because of the known differences between perception of natural speech as compared to the perception of synthetic speech. The findings of [Asakawa et al., 2003] revealed that the highest and most suitable rate for advanced users was always higher than for intermediate and novice users. Furthermore, the researchers found subjective and objective highest rates being almost equal for all subjects; the most suitable rate, however, was lower in objective than in subjective evaluation.

More recently, [Stent et al., 2011] conducted another experiment on the intelligibility of synthesized fast speech for the blind. Since a systematic comparison of the performance of different TTS systems for this user group was not available, the authors decided to run a pilot experiment on the intelligibility of several different TTS systems. In this, they opted for an open response recall task. The speech produced had a speaking rate of 300 to 550 words per minute which corresponded to approximately 1.5 times real time to 3 times real time. A significant effect of speaking rate was found, next to a certain influence of participant-related factors like age and familiarity with TTS. Important for the research conducted and presented in the current work, [Stent et al., 2011] noted blind users having a different performance metrics from sighted subjects and tending to prefer intelligibility over naturalness (cf. [Fellbaum, 1996], cf. also chapter 5). Moreover, also here visually impaired subjects showed a strong preference for one particular synthesizer and voice, in accordance with the assumptions made by [Nishimoto et al., 2006] and [Trouvain, 2007]. Results even revealed a main effect for synthesizer type. Similar observations will be discussed in relation to the perceptual evaluation of fast speech synthesized by means of unit selection speech synthesis described in chapter 8.2.1.

4.4 Summary and Conclusions

When implementing fast speech as a separate speaking style in unit selection speech synthesis, the perception and cognitive processing of fast speech is of main interest. Therefore, the current chapter examined different aspects of speech perception. The perception of natural fast speech in general was discussed first. In accordance with general explanations regarding speaking rate production and quantification outlined in chapter 2.1, first common models developed to describe and explain speech perception were described in section 4.1.1. Mechanisms of perceptional adjustment and compensation with regard to durational as well as spectral characteristics of fast speech were discussed subsequently. Remarks about natural fast speech perception were concluded with explanations on the units of speaking rate perception as well as the perception of speech rhythm in section 4.1.2.

In section 4.2, the perception of artificial fast speech was examined, distinguishing between time-compressed natural speech (cf. section 4.2.2) and synthesized (fast) speech (section 4.2.3). Both aspects are important for the perceptual evaluation of either the natural fast corpus recordings conducted with the selected speaker compared to time-compressed normal speech rate recordings presented in chapter 7.1.1, as well as regarding the evaluation of fast speech synthesized by means of different underlying corpora as described in chapter 8.2. Before going into detail, an overview was given over methods of perceptual evaluation of (synthetic) speech in section 4.2.1. The methods chosen for the perceptual evaluations conducted in connection with the work presented here were defined afterwards: In addition to judgments of intelligibility and naturalness for different sets of stimuli (cf. sections 6.2.2, 7.1.1, and 8.1), based on a Mean Opinion Score, the Word Error Rate was chosen to describe the perception of (ultra-) fast synthesized speech in chapter 4.2.3.

Possible differences between listener groups regarding the perception and judgment of synthesized fast speech were detailed in section 4.3. Based on the conclusion that individual listener characteristics may have a huge influence on judgments the experimental results of the evaluation of (ultra-)fast speech synthesized for this project will be analyzed with regard to individual pre-conditions of subjects and their experience with fast synthetic speech (cf. chapter 8.2). [Bradlow et al., 1995], for example, noted that familiarity with a speaker's voice led to an advantage in intelligibility. [Winters and Pisoni, 2004] investigated the extent to which expert knowledge of one form of synthetic speech may improve the perception of other forms of synthetic speech. Moreover, [Winters and Pisoni, 2004] stated that their research on the potential group of "expert listeners" may reveal what upper limits exist regarding the perception of ultra-fast speech. Also [Moos et al., 2008] wanted to determine if and how an extensive amount of listening experience influences the processing of fast speech. In connection with the research presented here, these aspects can unfortunately only be roughly touched in chapter 8.2.1.

Chapter 5

Preliminary Survey

The possibility to choose a fast speaking rate is reported to be essential for people who are reliant on artificial speech output like the blind and visually impaired ([Fellbaum, 1996], [Portele and Krämer, 1996], [Asakawa et al., 2002], [Chalamandaris et al., 2010], [McCarthy et al., 2013]). Hence, the goal of the work presented hereafter was to determine an optimal strategy for modeling fast speech in unit selection speech synthesis to provide potential users with a more natural sounding alternative for synthesized speech than the one provided by parametric synthesis. When preparing the empirical studies presented later on some fundamental questions came up: What do the blind and visually impaired really aim for concerning synthetic speech quality? Do they indeed prefer a monotonous fast speech synthesis being prosodically relatively close to natural fast speech as suggested by [Fellbaum, 1996]? Do they not mind a lack in naturalness at all, as long as acoustic transitions important for segment identification are adequately modeled as in formant synthesis [Moos and Trouvain, 2007]? What kind of speech quality do they prefer in general?

An early study about German speech synthesis applications for the blind and visually impaired was conducted by [Portele and Krämer, 1996]. The authors noted that *intelligibility* was the crucial factor for preferring certain speech applications over others, but the utterances generated by those applications often lacked the naturalness of human speech. Additionally, [Portele and Krämer, 1996] claimed that an adequate speech synthesis system should provide a speaking rate of at least three times the standard rate. However, this must not result in unintelligible speech output. For many users, the ease of use, flexibility, and robustness of a system were determined to be as important as speech quality in terms of naturalness. In a more recent study, [McCarthy et al., 2013] confirmed these findings. The researches investigated habituation to screen reader usage as well as switching behavior between different systems among visually impaired users in India. In their study, the loyalty to a certain system as opposed to the willingness to experiment with different systems across different user groups was investigated. Results suggested that for novice users of speech synthesis the main drivers of adoption of a specific screen reader software were a human sounding voice and voice quality in general, whereas the most important factors for advanced users were application support as well as the possibility to speed up the uttered speech to a certain extent. In this context, advanced users were found to be more comfortable with non-human sounding speech than novice users. As a conclusion, the authors suggested to aim for an "integrated" approach preserving a certain amount of naturalness of speech, and at the same time providing output at a speaking rate satisfying especially advanced users' needs with regard to desired speaking rate. This is in line with the findings of [Chalamandaris et al., 2010] who presented a unit selection TTS system optimized for usage in screen readers in Greek. They stated that TTS technology in general needed options for adaptation and customization in dedicated applications. In their investigations, screen reading software was expected to be able to cope with almost all kinds of text including English inclusions in Greek. Here, advanced users even suggested to provide an option allowing for degraded speech quality in exchange for increased speed.

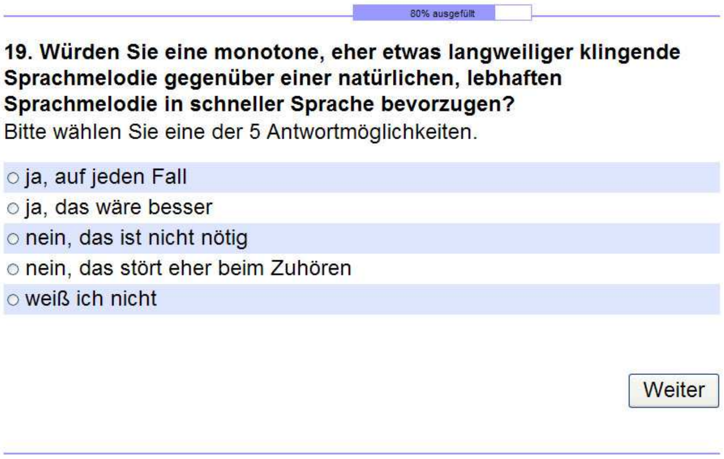
Since specific preferences of blind and visually impaired users have not been investigated for German TTS applications as much in detail as it would have been desirable for designing an optimal strategy for modeling fast (German) speech in unit selection speech synthesis it was decided to perform a preliminary survey among the prospective users before starting the main work on implementing fast speech in unit selection speech synthesis. Moreover, another goal of the preliminary survey was to not later on encounter the problem of "lack of understanding the users' needs" [Wagner, 2013]. The issues and results of the developed questionnaire are given in detail in the following sections (cf. [Moers et al., 2007]).

5.1 Methods

The questionnaire was designed to capture general as well as more fine-grained preferences of the target audience about the speaking style to be modeled. It was carried out online in an almost barrier-free environment. The questionnaire started with socio-demographic questions about the user's age, gender and ability for seeing. In the following, questions about the use of speech synthesis applications in general were asked. In particular, these questions were about

- the kind of assistive technologies in use (speech output, braille computer keyboard),
- the duration, regularity and amount of speech synthesis use,
- the fields of synthesis application (for private or business purposes),
- which particular system was preferred.

The subsequent part of the survey dealt with questions about the actual use of fast speech output as well as preferences concerning the tradeoff between



19. Würden Sie eine monotone, eher etwas langweiliger klingende Sprachmelodie gegenüber einer natürlichen, lebhaften Sprachmelodie in schneller Sprache bevorzugen?
Bitte wählen Sie eine der 5 Antwortmöglichkeiten.

○ ja, auf jeden Fall
○ ja, das wäre besser
○ nein, das ist nicht nötig
○ nein, das stört eher beim Zuhören
○ weiß ich nicht

Figure 5.1: Example of questionnaire item and possible answers.

naturalness, liveliness, and the possibility to have a synthesizer talk very fast. In detail, these questions asked about preferences regarding

- the possibility to choose fast speech output for the device in use,
- the use of fast speech rates in general,
- the sort of texts where fast speech was used most frequently (e.g. news, prose),
- the preferred intonation for fast speech (e.g. monotonous, lively),
- the distinctness of single phones,
- the realization of punctuation marks and pauses,
- the desirability of naturalness and,
- the willingness to pass on naturalness for the benefit of intelligibility.

Based on the findings of [Quené, 2007] who claimed that the most important parts of speech - which normally are content words (in contrast to function words) - are pronounced more clearly and a little bit slower than unimportant parts of speech, the last part of the survey included questions about the desirability of the distinction between content words and function words in terms of

- differences in speaking rate,
- differences in accentuation and,
- differences in intensity.

In total, the questionnaire comprised 23 questions [Moers et al., 2007]. Figure 5.1 gives an example of such an item. A detailed list of questions and possible answers can be found in appendix A.

5.2 Results and Discussion

Altogether, 100 blind or visually impaired subjects took part in the survey. 16 data sets were incomplete and therefore excluded from analysis. The remaining 84 subjects were between 18 and 70 years old (mean 37 years), 66 of them were male (78.6%). With regard to speech synthesis applications, 75% of the participants indicated to use speech output as part of a screen reader for both private and professional purposes. The most popular system was *JAWS Eloquence* ([FreedomScientific, 2011], cf. also [McCarthy et al., 2013]) which is a rule-based formant synthesizer providing synthetized speech output at a speaking rate of up to 23 syllables per second. It was used almost every day by 70% of the subjects. 95% of the subjects were using speech synthesis in general for more than 5 years [Moers et al., 2007]. This already hinted at a possible training effect likely to occur for this group of listeners (cf. chapter 8), and to be taken into account for investigations carried out later on.

60% of the respondents indicated to *always* change output rate to fast speech. Unfortunately, the exact amount of acceleration was not captured although it might have been advantageous to do so with regard to the perception experiments performed and discussed later on in chapter 8. [Quené, 1996] who generated texts with the "Apollo-spraaksynthese" for Dutch reported the highest intelligible speaking rate for a trained listener at approximately 8.64 syllables per second without any pauses included in the output; the most comfortable listening rate, however, was indicated at a speaking rate of 6.61 syllables per second (cf. also [Moers et al., 2007]). This is much less than what was required by [Portele and Krämer, 1996] as a "must" for an adequate speech output device for the blind and visually impaired. Moreover, these findings are also reflected in the results of the perceptual experiments presented in section 8.2.

Regarding the subjects' preferences concerning fast speech intonation and phrasing more than half of the subjects (51%) held that intonation had to be sustained, and even more respondents (65%) indicated that a monotonous intonation was neither desirable nor feasible. The claim of [Fellbaum, 1996] that blind and visually impaired users preferred a monotonous fast synthesis being prosodically relatively close to natural fast speech therefore seems to hold only in parts here and does not apply to this group of speech synthesis users in general.

For phrase boundaries, 96% of the subjects stated that the markedness of boundaries has to be preserved[1]. According to that, 81% of participants indicated preferring a full realization of pauses. Also, 87% of subjects voted for distinctness of single phones. The hypothesis that distinctness (or *intelligibility*) is the most important feature when using speech output devices is

[1] Since *JAWS Eloquence* allows for punctuation marks being explicitly announced it is not clear whether subjects voted for this feature being further available or for phonetic-prosodic markedness of boundaries by intonation.

supported here (cf. [Moers et al., 2007]). It was also observed by [McCarthy et al., 2013] and [Portele and Krämer, 1996]. Furthermore, the findings reported in section 8.2 strengthen this view as well.

A third of all participants judged *naturalness* as not being very important, and 25% would definitely do without it. In contrast, 40% indicated that they do not want to disregard naturalness completely (cf. [Moers et al., 2007]). This is in line with the findings of [McCarthy et al., 2013] again who stated that naturalness was not the most important factor for preferring a certain synthesis system, especially if the users were advanced in using assistive speech technology. In this context, some of the subjects also made use of the possibility to contact the investigator by email. The statements given below are examples of comments referring to the importance of naturalness and the disadvantages of concatenative speech synthesis in general.

> ”I am a synthesizer fan [formant synthesis; author's note]. I would rather pass on naturalness for the benefit of speech rate. Until now, all so called natural synthesis systems [concatenative synthesis; author's note] sound too staccato when adjusted to fast speech.”

> ”So called natural speech synthesis systems do actually not sound natural because up to now a natural flow [presumably referring to intonation and prosody; author's note] in speech could not be produced.”

> ”I prefer a synthetic speech synthesis to a natural speech synthesis system, although the latter indeed sounds more natural or melodious.”

These statements again showed that distortions of the speech signal occurring while using concatenative synthesis had such a huge negative effect on speech intelligibility that naturalness - although higher for voices used in this kind of speech synthesis systems - did not play a significant role for blind and visually impaired listeners. However, a considerable quantity of this group of speech synthesis users does not agree to disregard naturalness completely for the benefit of speaking rate (cf. [Moers et al., 2007]). The overall results of the preliminary survey thus encouraged the idea to investigate the possibility of generating fast speech in unit selection synthesis despite the disadvantages of this synthesis approach being clearly addressed.

Chapter 6

Speaker Selection and Evaluation

The continuous speech flow is accompanied by coarticulation and reduction, but nonetheless a sufficient contrast between neighboring segments is both necessary and achievable in successful human communication. According to Lindblom's H&H theory [Lindblom, 1996], a contrast is sufficient if it allows the listener to discriminate the signal to the extent necessary to identify the intended item in his mental lexicon. In contrast, the speaker produces speech earmarked and future-oriented. This causes a dilemma: On the one hand, the speaker tries to communicate with as little effort as possible. ”Hypospeech”, a less careful articulated speaking style, is the result of this economic constraint. On the other hand, the speaker wants to reach a communicative goal. Therefore, s/he needs to maintain phonetic contrasts necessary for comprehension. Thus, in situations where comprehension is more difficult (e.g. in a loud environment) or absolutely essential (e.g. when giving instructions) speakers tend to use ”hyperspeech”, a very exact and clear pronunciation style. For fast speech, one would normally expect speakers to use hypospeech - due to economy. However, speakers may be well able to speak both fast and clear (hyperspeech) within certain articulatory constraints if the situation requires it (cf. section 2.3).

Research in unit selection speech synthesis has shown that the quality of the speech synthesized for the most part is determined by the inventory speaker [Syrdal et al., 1997]. Skilled speakers who learned to speak with consistent voice quality and high articulatory precision over a long period of time will generally produce an inventory at higher quality and consistency than untrained speakers. If the inventory is based on fast speech the problem of articulatory precision and consistent voice quality will presumably increase. To create a useful fast speech inventory for the purpose of synthesizing fast speech in unit selection synthesis a suitable speaker had to be found who was able to produce the required speaking style in an optimal way. The procedure of speaker selection and evaluation is explained in detail in the following sections.

First, speaker requirements are derived from the characteristics of natural fast speech (cf. chapter 2.2) as well as from possible speaking strategies (cf.

chapter 2.3). Afterwards, the procedure of selecting an adequate speaker is described. In section 6.2.1, the acoustic characteristics of fast speech produced by the selected speaker while applying different speaking strategies are investigated and compared to the characteristics of the speaker's normal speech. It was expected to find significant differences in the acoustic characteristics of normal versus fast speech as well as between the two fast speech samples produced with different speaking strategies. Afterwards, a perceptual evaluation of the selected speaker's fast versus fast and clear speech is described and results are discussed (cf. section 6.2.2). The hypothesis was that listeners would be able to distinguish between *fast (and sloppy)* and *fast and clear* speech, and that they would judge the latter being more intelligible than the former.

6.1 Speaker Selection

Investigations into different speaking styles and the characteristics of fast speech have shown that *clear speech* is at an advantage regarding intelligibility compared to conversational speech both for normal and fast speaking rates. Clear speech was successfully elicited by repeated pronunciation of the same utterance by experienced speakers who had public speaking experience [Krause and Braida, 2002], [?]. Assuming that untrained speakers would reduce articulatory precision for the benefit of economic reasons to a greater extent than skilled speakers when speaking fast, the inventory speaker should fit the following criteria:

- S/he should be an experienced speaker who is able to speak both very fast and clearly. [Widera, 2000] pointed out that experienced speakers also showed a *more consistent* strategy for signaling vowel reduction levels which made correct perception of vowel realizations easier over time. Additionally, previous studies showed a maximum speaking rate at approximately eight syllables per second for German [Dellwo and Wagner, 2003] and Dutch [Janse, 2003a] when the speech was still highly intelligible. This rate was the set target for the fast speech inventory which was to be developed here.

- The speaking experience of the speaker should not emanate from a specific domain. S/he should not use a specific speaking style like news anchor or auction house style because such specific speaking styles are not transferable to other domains and therefore not suitable for usage in an open-domain unit selection speech synthesis system.

Beyond, based on the recordings made for the US English TIMIT corpus [Byrd, 1994] found that speech produced by male speakers was characterized by a greater spreading of phonological reduction than female speech. A year later, [Bradlow et al., 1995] confirmed that female speakers' speech was more intelligible than male speakers' speech (cf. [Altmann and Young, 1993], [Dupoux and Green, 1996]): The 4 talkers scored the most intelligible in their study were female, whereas the 4 talkers with the lowest intelligibility score were male (cf. also [Klatt and Klatt, 1990]; [Dupoux and Green, 1996]). However, the authors also stated that all of the male talkers showed a higher overall speaking rate than female talkers. The conclusion from their investigations was that the highest intelligibility could be reached by a

> "female who produces sentences with a relatively wide range in fundamental frequency, employs a relatively expanded vowel space that covers a broad range in F1, precisely articulates her point vowels, and has a high precision of inter-segmental timing." [Bradlow et al., 1995], p. 111.

In a more recent study of articulatory kinematics and precision in different speaking rate conditions, [Mefferd and Green, 2010] additionally stated that female talkers showed a significantly greater acoustic-phonetic specification in terms of the distance between two vowels in the vowel space than male speakers did. The authors related their finding to the smaller vocal tract size of females which may have articulatory precision made easier. In contrast, [Stent et al., 2011] noted that male voices outperformed female voices in their investigations. However, no significant main effect of the speaker's gender could be observed. [Trouvain et al., 2008] found that utterances produced with enlarged pitch range and a faster habitual articulation rate than the default setting of the diphone synthesis they investigated got better scores in the conducted evaluation, no matter whether they were performed by a female or a male speaker. Moreover, faster speakers appeared more competent and convincing, more confident, more intelligent, and more objective ([Smith et al., 1975] after [Trouvain et al., 2008]).

Based on these findings, the search for a suitable speaker started with a group of nine voluntary subjects. Six of the candidates were female and three male. Their age was between 21 and 34 years. All of them had either done corpus recordings for speech synthesis before or had other speaking related experiences as a radio presenter or similar. Test recordings were carried out in a sound-treated room. The speaking tasks included reading seven German sentences with differing length and content at a normal rate and repeated reading of the same sentences at a speaking rate as fast as possible. Afterwards, the recorded speech was judged by twelve phonetically trained listeners, mostly students at a high level of education or staff members. Judging criteria were

- the individual voice characteristics and appeal,
- the accuracy of articulation,
- the individual speaker's fastest possible speaking rate,
- the perceptual clarity regarding fast speech,
- the sustainment of voice quality and intensity, and
- the naturalness of intonation and pronunciation.

These criteria are considered the best guarantee for a high degree of naturalness in unit selection speech synthesis. The presumably most suitable speakers for a fast speech inventory, two female and one male speaker, were determined based on those judgments. After a second round of assessment which in addition included the factors

- kind of speaking experience and
- availability over time,

one of the female speakers turned out to be the most suitable candidate as she had done corpus recordings before and was available over a long period of time (cf. [Moers and Wagner, 2008], [Moers and Wagner, 2009]).

6.2 Speaker Evaluation

Since neither an acoustic nor a fine-grained perceptual analysis was done during the speaker selection procedure, again recordings at both normal and fast speech rate were carried out to verify that the selected speaker indeed was able to speak very clearly at the desired maximum speaking rate (eight syllables per second, cf. section 6.1). These recordings were based on a short text which was already used in the *BonnTempo-Corpus* [Dellwo et al., 2004] and was familiar to the speaker. It derived from the narrative *Selbs Betrug* by B. Schlink [Schlink, 1994] and included four main and three subclauses.

At first, the speaker was asked to read the text at a normal speaking rate in a speaking style suitable for usual corpus recordings, including consistent overall voice quality, consistent speaking rate, an accurate pronunciation and a neutral intonation. Afterwards, the speaker read the same text three times as fast as possible without taking special care of articulatory precision. Subsequently, another three fast recordings were conducted. Here, the speaker was asked to intentionally increase the articulatory effort and to produce fast speech as articulate as possible. In both the fast and the fast and clear condition, speech rate was intended to increase for each of the three repetitions. This way, a small corpus of recordings in fast and clear speech as opposed to normal as well as fast speech was obtained.

6.2.1 Acoustic evaluation

An acoustic analysis of the two different fast speaking styles was performed subsequently. The fast and fast and clear versions were compared to each other as well as to the normal rate version. The intention was to see how and to what extent the speaker avoided undesired effects like coarticulation and reduction in fast and clear speech reflected in acoustic measurements. It was expected that fast speech would contain more of those undesirable effects than the fast and clear utterances. Given that speaking rates for both the fast and the fast

version	number of pauses	mean pause duration
normal	24	788
fast	**13**	246^{+}
fast and clear	17	186^{+}

Table 6.1: Number of pauses and mean pause duration in milliseconds in normal, fast, and fast and clear speech. Smaller values for fast speech in bold. Significant duration differences between normal and fast as well as normal and fast and clear speech marked by plus.

and clear utterances were comparable (see below), the phenomena investigated here are not expected to derive primarily from the differences between normal and fast speech as discussed in chapter 2.2, but from the distinction between *conversational* and *clear* speech as referred to in chapter 2.3.

Recordings were labeled manually using the Praat software [Boersma and Weenink, 2010]. Because the amount of speech material available was not as substantial as it would have been desirable to perform an extensive acoustic analysis of both speaking styles produced at a fast rate compared to speech uttered at a normal speaking rate, only some acoustic characteristics were investigated. In detail, the following measures were analyzed:

- Pause number and duration
- Segment duration and elision
- Acoustic reduction of vocalic segments in terms of changes in formant frequencies and vowel space
- Acoustic reduction in terms of overall acoustic differences/spectral similarity between realizations

Pause number and duration

The most obvious characteristic to investigate was the number and duration of pauses. It was hypothesized that the number and duration of pauses decreased more in fast speech than in fast and clear speech compared to normal speech. Table 6.1 summarizes the findings.

It is shown that in fast speech the number of pauses indeed was somewhat smaller than in fast and clear speech. Nevertheless, this difference was not significant. All pause durations in fast and fast and clear speech, respectively, were normally distributed (Kolmogorov-Smirnov test, p>.05) and homogeneous in variances (Bartlett's test, p>.05). However, in contrast to our expectations the mean duration of pauses was longer in fast speech than in fast and clear speech. Also this difference was not significant for the two fast speaking styles. Only in view of normal speech, the differences in pause duration were significant for both fast and clear (Welch's t-test, t=4.4375, df=7.5118, p<.01) as well as for fast speech (Welch's t-test, t=4.1657, df=7.5761, p<.01) as was expected (marked by plus in table 6.1). Thus, the hypothesis that the number and duration of pauses decreased more in fast speech than in fast and clear speech compared to normal speech was true, but neither significant in terms of total number of pauses nor with regard to pause duration.

Segment duration

As pointed out in section 2.2.2, the duration of phones belonging to different phone classes decreases to various extents when speaking rate increases. Vocalic segments, for example, are the most elastic components of speech [Campbell and Isard, 1991], and are therefore expected to be shortened more than consonantal segments when articulated faster. On the other hand, if articulatory effort was increased to produce speech in a fast and well articulated manner, one could expect less shortening than in casually produced fast speech.

To find out to what extent the selected speaker decreased segment duration of phones belonging to different phone classes when producing fast compared to fast and clear speech, the recorded utterances were segmented manually. Afterwards, single and mean phone durations were computed for each of the recorded text sets separately by applying a Praat script (cf. [Boersma and Weenink, 2010]). Only existing segments were taken into account; segment elisions were examined in a separate step afterwards. The amount of shortening of segment durations in fast and fast and clear speech compared to speech uttered at a normal speaking rate was analyzed. It was calculated as the percentage of shortening from normal to fast or fast and clear speech, respectively. Results - separated for vocalic and consonantal segments - are listed in tables 6.2 and 6.3.

In contrast to our expectations, results indicate that segment durations in fast and clear speech most of the time were shortened slightly more than in fast speech compared to normal speaking rate. For vocalic segments, there were only two cases ([iː], [ɛ], cf. table 6.2, bold) where vowel duration for fast speech was smaller than for fast and clear speech. All singular vowel durations were normally distributed (Kolmogorov-Smirnov test, p>.05) and homogeneous in variances (Bartlett's test, p>.05). For all of them, differences in durations between fast and fast and clear speech were not significant. Since it was reported that tense vowels are shortened more than lax vowels when speaking rate increases (cf. chapter 2.2.3), it was also hypothesized that tense vowels - which in German can be seen to be identical to long vowels - would be shortened more in fast speech than in fast and clear speech compared to normal rate speech. To verify this hypothesis for the current data, vowel durations were grouped according to the tense-lax criterion. All durations of tense or lax vowels, respectively, were normally distributed (Kolmogorov-Smirnov test, p>.05). Durations of tense vowels were not homogeneous in variances (Bartlett's test, Bartlett's K-squared=3.9068, df=1, p<.05) whereas durations of lax vowels were. For both groups of vowels duration differences between fast and fast

Vowel	mean duration *(n)* [ms]	mean duration *(f)* [ms]	mean duration *(fc)* [ms]	shortening *(n)* to *(f)* (%)	shortening *(n)* to *(fc)* (%)
[aː]	107	57	56	47.1	47.7
[ɛː]	121	59^{+}	51^{+}	51.4	57.3
[iː]	82	**41^{+}**	51	**49.9**	38.5
[uː]	88	62	62	29.6	30.0
[øː]	151	74	71	50.9	52.7
[a]	78	54^{+}	49^{+*}	30.4	37.2
[ɛ]	96	**50^{+}**	53^{+}	**48.4**	44.4
[ɪ]	67	35^{+}	35^{+}	47.1	48.1
[ʊ]	61	42	41^{+}	31.0	32.5
[ʏ]	55	44	38	20.8	32.0
[ɒ]	123	**55^{+}**	57^{+}	**53.8**	53.5
[ə]	67	53	51	20.6	23.5

Table 6.2: Mean vowel duration in normal (*n*), fast (*f*) and fast and clear (*fc*) speech in milliseconds; *(n)* to *(f)* percentage of shortening and *(n)* to *(fc)* percentage of shortening. Smaller durations in fast speech in bold. Significant duration differences between normal and fast as well as normal and fast and clear speech marked by plus.

and clear speech were not significant. Nevertheless, the results showed that in both fast versions the shortening of long vowels (45.8% on average for fast and 45.2% on average for fast and clear speech) was considerably larger than the shortening of short vowels (35.5% on average for fast and 39.0% on average for fast and clear speech), as was expected. Interestingly, the statistical analysis revealed that the duration differences between long and short vowels *within* each sample were highly significant in the case of fast and clear speech (Welch's t-test, t=3.8933, df=56.327, p<.001), but just not significant in the case of fast speech (Welch's t-test, t=2.0062, df=48.952, p=.0504). This indicates that the durational contrast important for the differentiation between long and short vowels was sustained in fast and clear speech whereas it was not - or at least not to the same degree - in fast speech. These findings are in line with the observations made by [Mefferd and Green, 2010], and are further supported by the outcome of the acoustic analysis of vowel formant frequencies discussed in chapter 6.2.1.

As mentioned before, different classes of consonants are shortened in different ways and to different degrees when speaking rate increases. Therefore, the duration differences between fast and fast and clear speech were examined for single consonants first. Afterwards, consonantal segments were grouped by manner of articulation for further analysis. It appeared that there was

Consonant	mean duration *(n)* [ms]	mean duration *(f)* [ms]	mean duration *(fc)* [ms]	shortening *(n)* to *(f)* (%)	shortening *(n)* to *(fc)* (%)
[f]	126	69$^+$	65$^+$	55.1	58.7
[v]	91	38$^+$	33$^+$	58.5	63.5
[s]	99	57$^+$	48$^+$	42.9	51.3
[z]	93	**45**$^+$	46$^+$	**51.6**	50.4
[ʃ]	139	99	71	38.7	48.7
[ç]	126	55$^+$	55$^+$	56.1	56.9
[x]	86	46	41	47.1	52.0
[h]	70	**27**$^+$	29$^+$	**61.8**	58.6
[b]	73	41	39	44.3	46.1
[t]	97	39$^+$	39$^+$	59.4	59.8
[d]	52	**24**$^+$	25$^+$	**55.6**	53.7
[k]	91	**47**	51	**47.8**	43.5
[g]	84	**40**$^+$	44$^+$	**52.6**	47.6
[?]	47	16$^+$	18$^+$	64.6	60.7
[m]	108	**55**	58	**48.9**	45.7
[n]	86	**44**$^+$	44$^+$	**49.0**	48.3
[r]	71	**29**$^+$	31$^+$	**59.3**	55.8
[l]	71	44	43	37.8	39.1

Table 6.3: Mean consonant duration in normal (n), fast (f) and fast and clear (*fc*) speech in milliseconds; *(n)* to *(f)* percentage of shortening and *(n)* to *(fc)* percentage of shortening. Smaller durations in fast speech in bold. Significant duration differences between normal and fast as well as normal and fast and clear speech marked by plus.

no consistent pattern of change in duration when articulatory effort was increased in fast and clear speech compared to fast speech. All singular consonant durations were tested for normal distribution and homogeneity of variances. Results showed that durations of all consonantal segments were normally distributed beside for [d] (Kolmogorov-Smirnov test, D=.2496, p<.05). For consonants [k] (Bartlett's test, Bartlett's K-squared=4.2438, df=1, p<.05), [ʃ] (Bartlett's test, Bartlett's K-squared=4.6745, df=1, p<.05), and [t] (Bartlett's test, Bartlett's K-squared=6.8035, df=1, p<.01) durations were not homogeneous in variances. Across all consonantal segments, again almost half of the phonemes were shortened less in fast speech than in fast and clear speech (cf. table 6.3, bold). Duration differences between single consonants of fast versus fast and clear speech were only significant in the case of [ʃ] (Welch's t-test, t=-2.5712, df=6.089, p<.05). Although the overall duration differences for consonantal segments were not significant, the higher dispersion about the mean found for their duration distribution in fast speech as opposed to fast and clear speech potentially indicated that for the production of fast and clear speech a more consistent articulation strategy was used. This hypothesis will be investigated in more detail in section 6.2.1.

Looking at different groups of consonantal segments, various patterns of changes in duration were found. For plosives, for example, no regularity in shortening appeared at all: Some plosives were shortened more in fast than in fast and clear speech (e.g. [d], [k], [g]), others were not (e.g. [b], [t]). Fast speech plosives' durations were not normally distributed (Kolmogorov-Smirnov test, D=0.16425, p<.005) whereas they were in fast and clear speech. Plosives' durations were not homogeneous in variances (Bartlett's test, Bartlett's K-squared=7.6118, df=1, p<.01). The duration differences for plosives between the two fast speaking styles were not significant. For nasals, a more consistent pattern arose since both produced nasal segments [m] and [n] were shortened more in fast than in fast and clear speech. Nasals' durations were homogeneous in variances (Bartlett's test, p>.05) and normally distributed (Kolmogorov-Smirnov test, p>.05). Duration differences between fast and fast and clear speech were not significant again. In contrast, for fricatives the results of the duration analysis indicated that there was more shortening in fast and clear than in fast speech with the exception of [z] and [h]. Fricatives' durations were homogeneous in variances (Bartlett's test, p>.05) and normally distributed (Kolmogorov-Smirnov test, p>.05). The duration differences between fricatives emanating from fast and fast and clear speech, respectively, were significant (Welch's t-test, t=-2.1242, df=198.22, p<.05). Thus, fricatives were the only class of consonantal segments where a significant duration difference between fast and fast and clear speech was observable although the amount of shortening in fast and clear speech was higher than expected.

Segment elision

Assuming that fast speech was articulated less carefully than fast and clear speech it was also expected that for fast speech the number of elisions would be higher. According to [Koreman, 2006], this would result in a smaller realized speech rate for fast speech as opposed to fast and clear speech.

The vocalic segment estimated to be elided most frequently in both fast and fast and clear speech was the central vowel [ə], since in German it only occurs in unstressed syllables which tend to be reduced anyway. The same holds for the near-open central vowel [ɒ]. Because of the differing speaking styles it was expected that in fast speech [ə] and [ɒ] would be elided more often than in fast and clear speech. Both phonemes were counted as elided if no vowel-like structures were found in the signal. In total, 19 of 33 [ə] were left out in the fast version. For the fast and clear version, it was 18 (cf. table 6.4), so that it was concluded that there were no significant differences between fast and fast and clear speech in the number of [ə] elisions. Compared to normal speech, the amount of elisions however was significant for both fast speech versions (fast speech: χ^2=7.6809, df=1, p<.01; fast and clear speech: χ^2=6.75, df=1, p<.01). For [ɒ], none were elided in fast and clear speech, and only one out of 15 was left out in fast speech. This difference in number of [ɒ] elisions was only marginal and did not show enough evidence to argue for a definable difference between the two fast speaking styles.

Looking at consonantal segments, it was found that [?], [t], and [s] were elided more frequently in the fast speech versions. The number of elisions of those segments in fast and fast and clear speech in contrast to normal speech is summarized in table 6.4. Again, differences in the amount of segment elisions between fast and fast and clear speech compared to normal speech were not significant beside for the glottal stop [?] (χ^2=8.7273, df=1, p<.01 in fast speech, and χ^2=5.0845, df=1, p<.05 in fast and clear speech). Other consonantal segments were rarely or never elided in both fast speech versions compared to speech produced at normal speaking rate. However, looking at the total number of elided segments in fast speech compared to normal speech, there is a significant difference observable (χ^2=9.8, df=1, p<.01), whereas the amount of elisions in fast and clear speech compared to normal speech is not significant. Coming back to [Koreman, 2006], this result indicates a significantly lower realized speech rate for fast speech than for fast and clear speech compared to the intended speech rate. The ratio of realized to intended speech rate in terms of number of segments consequently points to a lower "Articulatory Precision Index" ([Koreman, 2006]) for fast speech compared to fast and clear speech as outlined in chapter 2.1. Additionally, it is remarkable that two of the segments with a noticeable amount of elisions ([t], [s]) in fast as well as in fast and clear speech were alveolar articulated phonemes. This result leads to the assumption that the tongue movement towards the alveolar ridge to produce the closure or constriction for the respective segment was probably not completely realized anymore when producing fast or fast and clear speech,

segment	overall number normal speech	overall number fast speech	overall number fast and clear speech
[ə]	33	**14**$^+$	15$^+$
[ɒ]	15	**14**	15
[ʔ]	45	**21**$^+$	26$^+$
[d]	39	**37**	39
[h]	12	**9**	12
[s]	30	**27**	30
[t]	60	**49**	60

Table 6.4: Segment elisions for fast and fast and clear speech compared to normal speech rate in overall number of segments. Smaller values for fast speech in bold. Significant differences between normal and fast as well as normal and fast and clear speech marked by plus.

respectively, though the movement was not affected to the same extent for the two fast speaking styles (cf. [Moers and Wagner, 2008], [Moers and Wagner, 2009]).

The presented analysis of durational differences for vocalic and consonantal segments showed no consistent, significant differences between fast and fast and clear speech. It was concluded that the measured durational differences of segments did not provide enough evidence to reliably predict the ability of the selected speaker to produce two different speaking styles when speaking fast. Only the total amount of segment elisions showed a clear tendency to omit more segments in fast speech as opposed to fast and clear speech. As this was a very global observation, other possible sources of variation were investigated in the next step.

Acoustic vowel reduction

Along with their higher elasticity concerning durational changes, vowels can also be more reduced in terms of other acoustic measurements than consonantal segments when speaking rate increases [Campbell and Isard, 1991]. The acoustic characteristics important for the identification of a vowel are - aside from the target formant frequencies - especially the formant transitions [Amano-Kusumoto and Hosom, 2010]. This allows a speaker to mainly shorten the inner part of a vowel where formant frequencies are regarded to stay relatively stable when speaking faster. When comparing fast as well as fast and clear speech to speech articulated at a normal speech rate it was expected that the reduction of formant frequencies would be stronger in fast speech than in fast and clear speech.

Vowel	f1 *(n)* [Hz]	f2 *(n)* [Hz]	f1 *(f)* [Hz]	f2 *(f)* [Hz]	f1 *(fc)* [Hz]	f2 *(fc)* [Hz]
[aː]	964	1957	743	2047	760	2003
[ɛː]	748	2514	395	2211	479	2223
[iː]	288	2630	341	2200^{+}	364	2272^{+}
[uː]	356	1531	409	1664	451	1700
[øː]	347	1988	389	1844	374	1891
[a]	904	2022	800	2147	774^{+}	2036
[ɛ]	639	2166	589	1903^{+*}	634	2027^{+*}
[ɪ]	438	2437	399	1989^{+*}	430	2086^{+*}
[ʊ]	459	1678	598	1780	601^{+}	1914
[ʏ]	577	2760	415	1900^{+}	437	1985^{+}
[ə]	410	1926	588	1877	554	1907
[ɒ]	665	1765	552	1849	694	1953

Table 6.5: Mean first (F1) and second (F2) formant frequencies of distinct vowels in normal (*n*), fast (*f*) and fast and clear (*fc*) speech in Hertz. Significant differences between normal and fast as well as normal and fast and clear speech marked by plus. Significant differences between fast and fast and clear speech marked by asterisk.

To find out to what extent the selected speaker actually acoustically reduced vowel formant frequencies when speaking fast, and to compare the two different fast versions to the normal speech version, the first and second formant of each segmented vowel were computed for each of the three recorded text sets per speaking style by applying a Praat script (cf. [Boersma and Weenink, 2010]). The script went through the sound files and corresponding TextGrid files in a given directory, opened each pair of sound and TextGrid, and calculated the formant values at the midpoint of each labeled vowel interval. Measurements are summarized in table 6.5. All singular formant frequencies were tested for normal distribution and homogeneity of variances. Results showed that all formant frequencies were normally distributed beside the first formant F1 for [a] in fast speech (Kolmogorov-Smirnov test, D=.29101, p<.05) and [ɪ] in fast and clear speech (Kolmogorov-Smirnov test, D=.29816, p<.05). Formant frequencies of F1 for [a] (Bartlett's test, Bartlett's K-squared=17.54, df=1, p«.001), [ʏ] (Bartlett's test, Bartlett's K-squared=7.5057, df=1, p<.01), [iː] (Bartlett's test, Bartlett's K-squared=5.3615, df=1, p<.05), and [ɒ] (Bartlett's test, Bartlett's K-squared=11.272, df=1, p<.001) were not homogeneous in variances. Differences between single formant frequencies of fast versus fast and clear speech were not significant beside for the second formant F2 for [ɪ] (Welch's t-test, t=2.0675, df=42.959, p<.05) and [ɛ] (Welch's t-test, t=3.6061, df=45.856, p<.001), cf. table 6.5, asterisk).

As can be derived from the results shown in table 6.5, the formant fre-

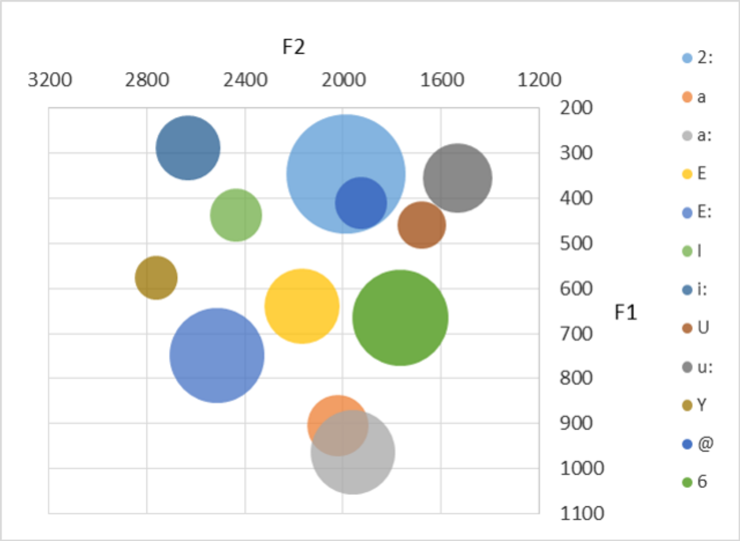


Figure 6.1: Vowel chart for normal speech

quency reduction found for fast as well as fast and clear speech compared to normal speech mainly affected the second formant. These findings led to the conclusion that vowel articulation was more centralized in terms of a more retracted tongue position, but not pointing into the direction of the neutral central vowel [ə]. This is in line with the previous finding that alveolar consonant articulation is not executed to the same degree anymore, interpreted to indicate that the tongue movement towards the teeth ridge was not realized to the same extent when producing fast or fast and clear speech compared to normal speech. In contrast to our expectations, the results also show that formant frequencies were not reduced significantly more in fast speech compared to fast and clear speech as well as opposed to speech produced at a normal speaking rate. There were only four out of twelve cases where the second formant F2 in both fast and fast and clear speech was significantly lower than the one in normal speech ([iː] in fast speech: Welch's t-test, t=5.0795, df=9.8316, p<.001, and [iː] in fast and clear speech: Welch's t-test, t=5.0805, df=7.1141, p<.01; [ɛ] in fast speech: Welch's t-test, t=5.8535, df=14.305, p«.001, and [ɛ] in fast and clear speech: Welch's t-test, t=2.9998, df=15.518, p<.01; [ɪ] in fast speech: Welch's t-test, t=7.8876, df=19.603, p«.001, and [ɪ] in fast and clear speech: Welch's t-test, t=6.0308, df=19.591, p«.001; [ʏ] in fast speech: Welch's t-test, t=5.6042, df=5.5479, p<.005, and [ʏ] in fast and clear speech: Welch's t-test, t=7.3072, df=2.7534, p<.005), indicating a more retracted articulation again. The first formant F1 was only affected in two cases ([a]: Welch's t-test, t=2.5338, df=18.573, p<.05, [ʊ]: Welch's t-test, t=-2.8334, df=13.941, p<.05) in fast and clear speech; cf. table 6.5, marked by plus), not revealing any regular pattern of changes in formant frequencies in fast or fast and clear speech versus normal speech.

To visualize the differences between vowel spaces in terms of vowel formant frequencies and duration in normal speech versus fast and fast and clear speech, the first and second formant frequency values were transferred into a Cartesian coordinate system with the horizontal axis showing the second formant F2 reflecting the tongue position (”backness”), and the first formant F1 on the vertical axis reflecting the tongue height (”degree of opening”). The zero intercept point of the two axes was placed at the upper right corner to reflect the vowel space in accordance with the IPA vowel chart[1]. Each circle represents a distinct vowel. The center of the respective circle displays the mean first and second formant frequency, whereas the size of the area of the circle reflects the relative mean duration of the vowel.

Comparing figure 6.1 to figures 6.2 and 6.3, it becomes immediately evident that the vowel space for both fast speech versions is much smaller than the one for speech articulated at normal rate. Additionally, from the relative size of the circles which reflect the relative mean duration of the specific vowel it becomes clear that durational differences between long and short vowels are more distinct in normal speech (cf. section 6.2.1). One can also see that the circles in figure 6.2 displaying the characteristics of vowels produced in fast speech are showing more overlap than the circles in figure 6.3 representing distinct vowels produced in fast and clear speech. Those findings support the hypothesis that fast speech was articulated less precisely and distinctively, and thus with more categorical overlap than fast and clear speech. To verify these assumptions, the vowel space area and the vowel space dispersion in terms of the Euclidean distances to the hypothesized center of the speaker's vowel space were calculated. According to [Bradlow et al., 1995] and [Maniwa et al., 2009], a more clear articulation would be accompanied by an expanded vowel space and a higher vowel space dispersion compared to a more conversational speaking style.

To determine the vowel space area of the respective speaking styles produced by the selected speaker, as a first step the three vowels encompassing most of the vowel space across the different speaking styles were selected. Those vowels were [iː], [aː], and [uː]. Average formant frequencies for F1 and F2 were derived from the results of the acoustic analysis listed in table 6.5. The triangular vowel space area was then calculated by using Heron's formula and the Pythagorean theorem (cf. [Jacewicz et al., 2007]):

Area = SQRT(s(s-a)(s-b)(s-c))
where s = (a+b+c)/2
and $a^2+b^2=c^2$

Results showed that in contrast to our expectation the triangular vowel space area for fast and clear speech (101629,38 Hz^2) was slightly smaller than

[1] https://www.internationalphoneticassociation.org/content/ipa-vowels, last visited August 14, 2015

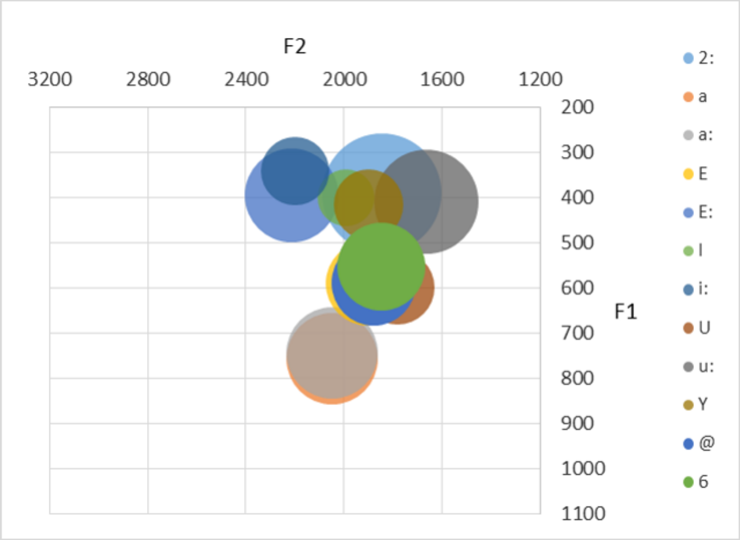


Figure 6.2: Vowel chart for fast speech

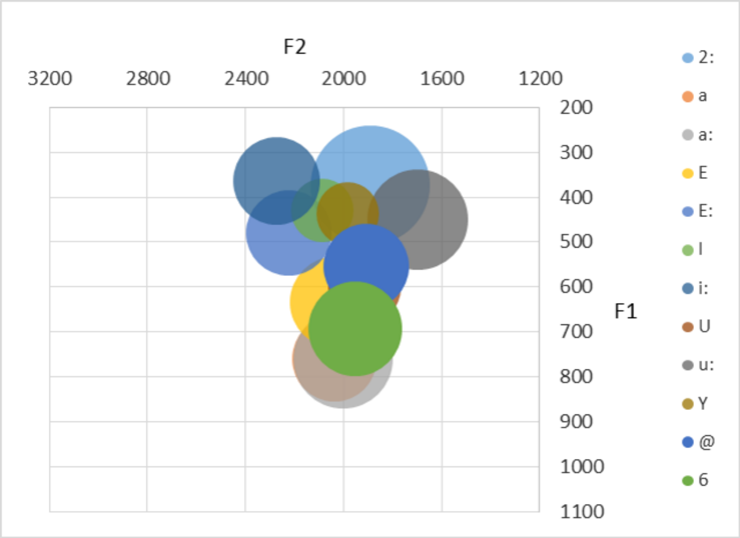


Figure 6.3: Vowel chart for fast clear speech

the triangular vowel space area for fast speech (102694,56 Hz^2). However, this difference between vowel space areas was very small with a discrepancy of approximately 1% in total only. Therefore, the hypothesis that fast and clear speech would show a larger vowel space area than fast speech had to be rejected. Compared to the triangular vowel space for normal speech, which spanned 348414,48 Hz^2, vowel space areas for both fast speech versions were considered to be significantly smaller, as was expected. All in all, the analysis of the vowel space areas of the different fast speaking styles failed to show a significant difference between fast versus fast and clear speech as well.

According to [Bradlow et al., 1995], the vowel space dispersion measured by Euclidean distances to the hypothetical vowel space center of the speech produced by a specific speaker might give a clearer picture of how individual vowels are distributed across the vowel space as it is calculated based on individual vowel tokens and not - as the vowel space area - based on average formant frequency values for certain vowels. As mentioned earlier, figures 6.2 and 6.3 give the impression that vowels produced in fast and clear speech were more distinct than vowels produced in fast speech. Therefore, it was decided to also analyze the vowel space dispersion for the two different fast speaking styles. It was expected to find a more dispersed vowel space for fast and clear speech than for fast speech since acoustic differences between vowels may have been kept more distinct and more stable, as established by higher Euclidean distances from the hypothetical center of the speaker's vowel space. First, Euclidean distances were calculated for single vowels deriving from the two fast speech versions separately; results are shown in table 6.6. Afterwards, the overall vowel space dispersion of the different fast speech versions was compared to the overall vowel space dispersion of speech produced at normal speaking rate. Looking at the mean Euclidean distances listed in table 6.6, it becomes obvious that there is no clear tendency for vowels to be more dispersed in fast and clear speech than in fast speech. After testing the Euclidean distances for all single vowels for normal distribution (Kolmogorov-Smirnov test, $p>.05$ for all vowels) and homogeneity in variances (Bartlett's test, $p>.05$ for all vowels), a series of Welch's t-tests did not show any significant differences in Euclidean distances to the hypothetical vowel space center between fast and fast and clear speech ($p>.05$ for all vowels).

Subsequently, the Euclidean distances calculated for the different speaking styles as a whole were tested for normal distribution and homogeneity of variances. Results showed that only the Euclidean distances for speech produced at a normal speech tempo were normally distributed (Kolmogorov-Smirnov test, $p>.05$); Euclidean distances for fast (Kolmogorov-Smirnov test, $D=.13167$, $p<.05$) as well as for fast and clear speech (Kolmogorov-Smirnov test, $D=.12472$, $p<.05$) were positively skewed, potentially indicating a more centralized articulation for these speaking styles. Overall Euclidean distances were homogeneous in variances (Bartlett's test, $p>.05$) for all speaking styles. When compared to the overall vowel space dispersion for speech produced at

vowel	Euclidean distance in fast speech in Hz	Euclidean distance in fast and clear speech in Hz
[ə]	259	214
[ɒ]	**211**	307
[øː]	221	200
[a]	**423**	438
[aː]	476	387
[ɛ]	**159**	179
[ɛː]	277	259
[ɪ]	**234**	250
[iː]	**310**	366
[ʊ]	608	582
[uː]	**370**	402
[ʏ]	294	191

Table 6.6: Mean Euclidean distances to hypothetical vowel space center for fast and fast and clear speech in Hz. Smaller values for fast speech in bold.

normal rate, dispersion for fast speech (Welch's t-tests, t=-3.7171, df=104.77, p=.0003) as well as for fast and clear speech was found to be significantly smaller (Welch's t-tests, t=-3.702, df=106.4, p=.0003). However, it was concluded that also the investigation of the vowel space dispersion did not give enough insight into the acoustic differences between fast and fast and clear speech to decide whether the selected speaker was able to produce fast and clear speech when enhancing articulatory effort. Because the results of the acoustic analyses conducted and described previously did not reveal any significant differences between fast and fast and clear speech, it was assumed that single measurable acoustic features were not the key to capture the difference between the two fast speaking styles. As the differences seemed to be much more fine-grained, another approach of analyzing the collected data was chosen subsequently which is examined in the follwoing section.

Spectral similarity

Since the results of the analysis of static acoustic features did not allow for a definite conclusion about the speaker's ability to produce two distinguishable speaking styles when speaking fast, a method developed by [Wade et al., 2010] and [Lewandowski, 2011] was applied to investigate the *spectral similarity* between the two fast speaking styles by comparing amplitude profiles. Following [Wade et al., 2010], amplitude profiles are to be seen as

> "representations that more faithfully encode the speech signal as it unfolds over time without making specific assumptions about what

> types of cues might be extracted or which regions of the signal are the most important.” [Wade et al., 2010], p. 10.

[Lewandowski, 2011] underlines that this method does not give a picture of momentary static features of the speech signal, but rather allows the user to compare similar speech signals by taking into account dynamic features revealing important additional information. The method originally was developed to compare word pairs by cross-correlating amplitude envelopes of the two words of a pair. Amplitude envelopes are computed for four frequency bands equally spaced on a logarithmic scale ranging form 80 to 7800 Hz, using a sampling rate of 500 Hz. The amplitude envelopes of the two words were then cross-correlated separately, and the highest resulting value was taken as a similarity measure. The similarity score also took distortions in the temporal domain as one of several dimensions of acoustic distance into account. This approach has shown to be useful when investigating acoustic differences assumed to be very fine-grained (cf. [Samlowski et al., 2013]).

Applying this method, the similarity scores of pairs of nine short phrases extracted from the recordings of the two different fast speaking styles were investigated. The excerpts were expected to show both coarticulation and reduction effects. Extracted phrases are listed in appendix B. Based on the assumption made previously that fast and clear speech was articulated more consistently and precisely, it was hypothesized that similarity scores for pairs of excerpts from the fast and clear speech would be higher than the similarity scores for pairs of excerpts from fast speech. A Wilcoxon rank sum test showed that this was not the case; similarity scores for fast speech excerpts were not significantly different from similarity scores for fast and clear speech excerpts. It was concluded that fast and clear speech was not produced with a higher consistency than fast speech.

Since similarity scores within each group of excerpts (that is: excerpts from recordings of the same speaking style) were quite high, it was further assumed that similarity scores for pairs of excerpts from recordings of different speaking styles had to be significantly lower then similarity scores for pairs of excerpts deriving from recordings of the same speaking style if each group represented a distinct speaking style. Indeed, significant differences were found for similarity scores of pairs of excerpts from fast and similarity scores of pairs of excerpts from fast and clear speech when compared to similarity scores for mixed pairs (Wilcoxon rank sum test; W=2610, p=.0021 for fast and clear speech; W=2384.5, p=.0424 for fast speech).

This way, it was eventually verified that the two speaking styles produced at a fast speaking rate were significantly different from each other. However, it was still not possible to draw the conclusion that fast and clear speech was indeed articulated more precisely compared to fast speech although the level of significance of the difference between similarity scores of pairs of excerpts from fast and clear speech compared to similarity scores for mixed pairs was much higher than the level of significance of the difference between similarity

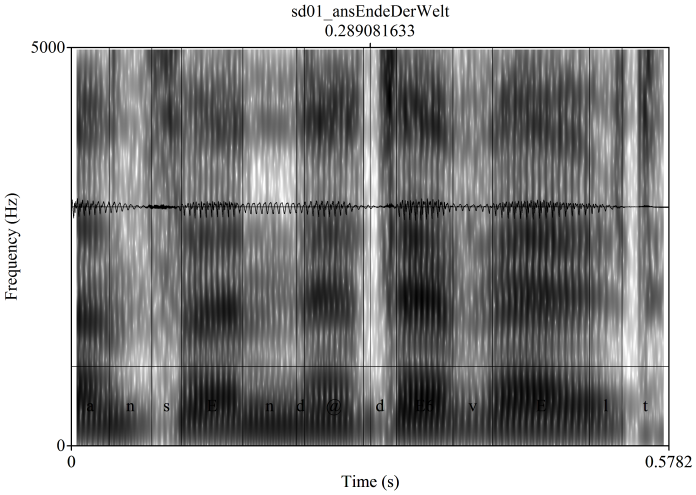


Figure 6.4: Spectrogram showing the excerpt ”ans Ende der Welt” (to the end of the world) of a fast and clear speech version.

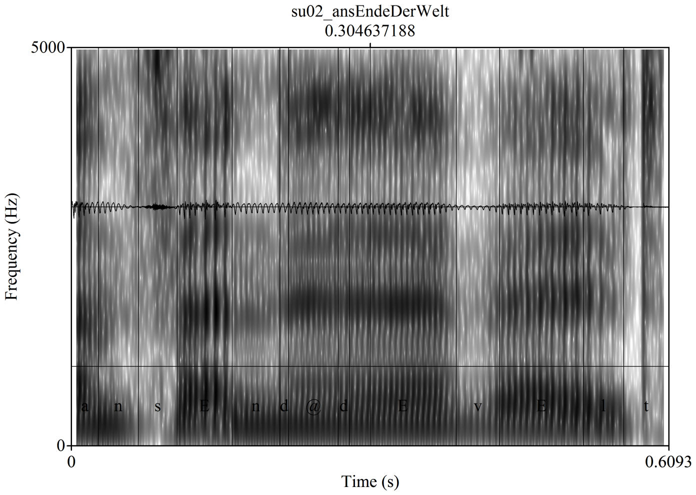


Figure 6.5: Spectrogram showing the excerpt ”ans Ende der Welt” (to the end of the world) of a fast speech version.

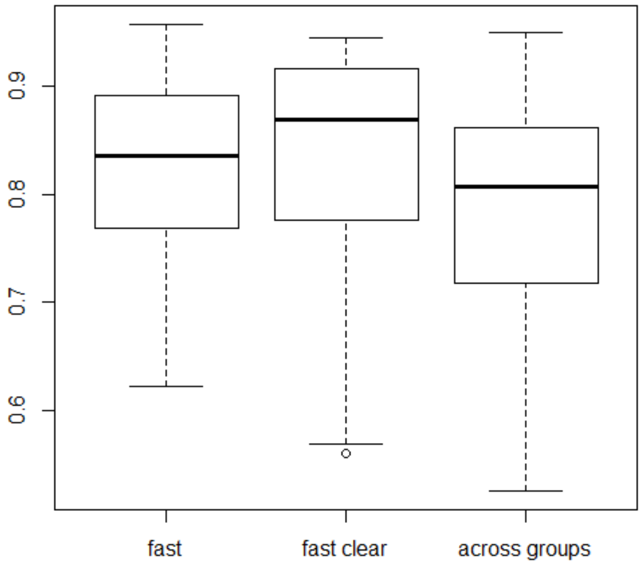


Figure 6.6: Similarity scores for excerpts derived from fast speech, fast and clear speech, and across groups.

scores of pairs of excerpts from fast speech compared to similarity scores for mixed pairs. Therefore, it was decided to additionally perform a perceptual evaluation of the acoustically analyzed excerpts of fast versus fast and clear speech with the aim to substantiate the assumption that the selected speaker was indeed able to produce two different, perceptually distinguishable speaking styles when speaking fast.

6.2.2 Perceptual evaluation

As the speaker was selected to create a fast-and-clear-speech inventory to be used in unit selection speech synthesis, it was important to evaluate whether the lack of significant measurable acoustic differences found previously would also become apparent in a perception experiment. If the speaker was able to produce fast speech more clearly by enhancing the articulatory effort, listeners would prefer the fast and clear utterances over the fast utterances with regard to intelligibility [Krause and Braida, 2002], [Adank and Janse, 2009]. If so, fast and clear speech produced by the selected speaker would fulfil the defined criteria to create an applicable fast-and-clear-speech unit selection inventory.

Methods

For the perception experiment, the same nine excerpts of short phrases from fast and fast and clear speech investigated acoustically earlier were chosen as stimuli. The perception experiment was created consisting of nine subsets, each of them containing 15 pairs of identical excerpts (with regard to contents) deriving from the different fast speaking styles as stimuli (cf. [Pickett and Pollack, 1963]). It was implemented by using the Praat ExperimentMFC environment [Boersma and Weenink, 2010] and conducted in a quiet environment. Stimuli were presented via earphones. The order of stimuli within each group of pairs was mixed randomly. The nine excerpts had been selected such that the linguistic content was still intelligible. However, to avoid problems in comprehension the text of each excerpt was displayed at the beginning of each subset. Furthermore, each stimulus pair could be replayed up to three times. Altogether, subjects were presented with 135 stimuli. They were instructed to indicate from each pair the realization which was more intelligible. The more intelligible version was credited one point, the sum of all points was interpreted as the respective "intelligibility score". 23 participants judged the presented pairs. It was expected that the fast and clear utterances would get a higher intelligibility score than the fast speech utterances (cf. [Moers and Wagner, 2008], [Moers and Wagner, 2009]).

Results

Because of the varying intended speech rates of the three different versions (cf. chapter 6.2) for fast and fast and clear speech, respectively, the exact speaking rate of each single version was estimated in syllables per second (cf. table 6.7, first column). Note that the order of the versions with regard to speaking rate does not reflect the order of production although in production it was intended to increase speaking rate from one sample to the next. However, there were only slight differences in speaking rate within and between groups which were not significant. Thus, different versions were still comparable.

In a second step, the arithmetic mean of the speaking rate was calculated for each pair of most similar fast versions. In order to account for the varying speaking rate within each pair, the respective intelligibility score of each involved version was divided by the exact speech rate and then multiplied by the mean value of the pair it belonged to. This way, a normalized value was obtained which gave an account of the intelligibility score relative to the speaking rate. Figure 6.7 shows that for all pairs of both speaking style groups, the fast and clearly articulated versions performed significantly better than the fast versions. A chi-square test confirmed these findings (χ^2, df=196, p<.001).

By means of the perceptual evaluation it was finally verified that the selected speaker indeed was capable of producing fast speech more clearly - and therefore more intelligibly - by enhancing articulatory effort. Thus, she was qualified to create a fast speech inventory articulated as accurately as possible

version	speech rate	mean speech rate	intelligibility score	normalized intelligibility score
sd01	7.53	7.69	701	716.21
sd02	8.26	8.32	576	580.16
sd03	7.25	7.30	670	674.76
mean/total	**7.68**		**1947**	
su01	7.35	7.30	568	563.96
su02	8.38	8.32	247	245.24
su03	7.85	7.69	342	334.89
mean/total	**7.86**		**1157**	

Table 6.7: Speaking rate of fast and fast and clear versions in syllables per second, mean speech rate for similar pairs, intelligibility score, and normalized intelligibility score.

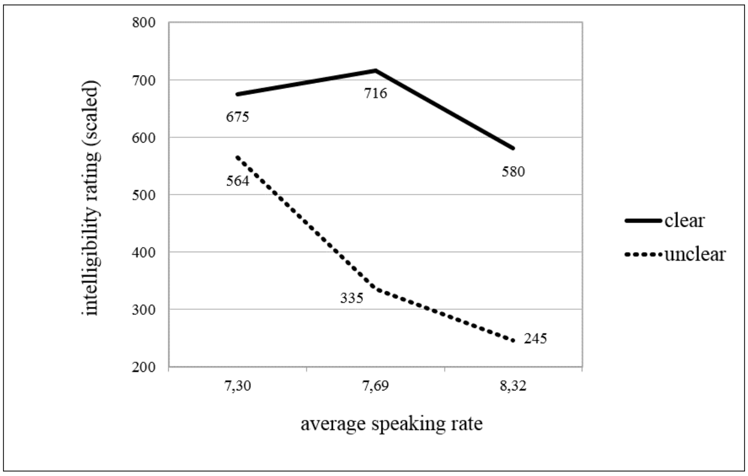


Figure 6.7: Normalized intelligibility scores for fast (dotted line) versus fast and clear (solid line) versions.

suitable for usage in unit selection speech synthesis (cf. [Moers and Wagner, 2008], [Moers and Wagner, 2009]).

6.3 Summary and Conclusions

After selecting a speaker who seemed to be able to reduce most of the undesirable phenomena usually occurring in fast speech to the required amount by enhancing the articulatory effort during fast speech production, an analysis of acoustic characteristics of two different fast speaking styles, fast as opposed to fast and clear, produced by the selected speaker was conducted. Concerning phrasing and pausing, it became apparent that in fast speech the total number of pauses decreased compared to fast and clear speech, but not to a significant degree. For vowels, segment duration and formant frequency analysis showed almost no significant differences between fast and fast and clear speech. Only systematic duration differences between tense and lax vowels were sustained in fast and clear speech to a significantly higher degree than in fast speech. Even a more detailed investigation of vowel characteristics in terms of vowel space area and vowel space dispersion analysis did not reveal the desired significant results. For consonantal segments, segment durations did not show a significant difference between the two fast speaking styles either. Just the total amount of elided segments, vowels as well as consonants, was significantly higher in fast speech than in fast and clear speech. Only by comparing spectral similarity for several excerpts from the different fast speaking styles it was finally possible to substantiate that the two speaking styles produced were indeed significantly different from each other in terms of acoustic characteristics. However, a conclusion about the difference in intelligibility of both speaking styles was not deducible. Therefore, a perceptual evaluation of the excerpts analyzed acoustically earlier was carried out to confirm the suitability of the selected speaker to create a fast speech inventory to be used in unit selection speech synthesis on the basis of the listeners' perception. Results showed that listeners clearly preferred stimuli containing excerpts of the fast and clear speech to stimuli consisting of excerpts from fast speech with regard to intelligibility. This way, it was verified that the selected speaker was indeed able to produce the required speaking style in an optimal way and therefore was suitable to create a fast speech inventory to be used in unit selection speech synthesis.

Chapter 7

Implementing Fast Speech in Unit Selection Synthesis

As discussed in chapter 3, there are several options to model fast speech in speech synthesis. The first one is to linearly accelerate normal speech by means of duration manipulation. The generated output often shows artifacts known to appear when using algorithms such as TD-PSOLA [Moulines and Charpentier, 1990], [Liu and Zeng, 2006], and does not sound very natural. The second option is to mimic certain prosodic features typical for fast speech such as fewer and shorter pauses or decreased strength and number of prosodic boundaries. Previous studies indicate that this approach leads to decreased intelligibility of fast speech [Portele, 1997], [Brinckmann and Trouvain, 2003], [Janse, 2003b], [Adank and Janse, 2009]. A clear pronunciation is preferred over synthesized speech which shows typical phonetic characteristics of natural fast speech. Therefore, the approach chosen here included the creation of an independent unit selection inventory for fast speech inherently showing segmental and supra-segmental characteristics of natural fast speech to enhance naturalness. At the same time, too heavy reduction and coarticulation typical for natural fast speech produced without any additional articulatory effort had to be avoided as much as possible for the benefit of intelligibility.

After confirming the selected speaker indeed was able to produce speech both fast and clearly articulated, a fast speech unit selection corpus to be implemented in the BOSS unit selection speech synthesis system (cf. section 3.1.2) was created. The development of the fast speech corpus as well as of a parallel corpus in normal speech rate, recorded and prepared for the purpose of comparison, is described in detail in the following sections. First, the recording procedure is outlined in chapter 7.1. To investigate the intelligibility and naturalness of recordings made at different speaking rates, the same utterances deviating from either the normal or the fast speech corpus were manipulated with regard to their duration, and perceptually evaluated afterwards. It was anticipated that clearly articulated natural fast speech would be as intelligible as natural normal speech when accelerated to the same (faster) speaking rate.

At the same time, utterances based on fast speech were expected to have an advantage in terms of naturalness as duration manipulation had to be less extensive than for normal speech recordings to generate a pre-defined (ultra-) fast speaking rate.

In the next section, the processing steps necessary to prepare a unit selection corpus are examined in more detail. The preparation of the unit selection inventory is one of the most time consuming steps during the development of new voices or speaking styles for unit selection synthesis, as usually a lot of manual labeling is required. Therefore, to label speech in normal speech tempo automatic labeling techniques are preferred. However, since the quality of the synthesized speech largely depends on "label timing accuracy" (LTA, [Kominek et al., 2003]), using the same segmentation algorithm for both normal and fast speech corpus recordings might result in a considerably increased amount of incorrect labels for fast speech utterances. If so, automatic phone segmentation would not be applicable to fast speech corpus recordings, even if the fast speech was articulated as accurately as possible. In section 7.2.1, the results of an automatic segmentation of the fast speech corpus recordings are outlined. Processing steps included the adaptation of already existing transcriptions to the needs of the BOSS synthesis system, the automatic segmentation of the corpus recordings by means of an HTK-based aligner adapted to German ([Dragon, 2005]), and subsequently the analysis of the label timing accuracy for both corpora.

Another important prosodic factor in the production of natural sounding synthetic speech is the accurate prediction of the duration of phonetic segments [Carlson et al., 1979]. Considering the results of [Janse, 2003b], it was decided to create segment duration prediction models by building CART-based regression trees [Breiman et al., 1984] for the labeled normal and fast speech corpus recordings separately, taking into account important phonetic and prosodic features influencing segmental duration. It was hypothesized that the generated duration prediction models showed significantly higher correlations between observed and predicted durations with normal speech rate utterances than with fast speech rate utterances because of the higher amount of coarticulation and reduction phenomena expected to occur in the latter despite maximized articulatory precision. Results of a comparative analysis of the generated CART-based duration prediction models for both corpora are presented in section 7.2.2.

7.1 Corpus Recordings

Text materials for corpus recordings consisted of 400 sentences which were selected randomly from the BITS Corpus for German [Schiel et al., 2006]. The BITS Corpus was chosen because of its phonologically balanced design meeting the general criteria of unit selection speech synthesis systems [van Santen and Buchsbaum, 1997], [Bozkurt et al., 2003]. It was developed especially for

German diphone and unit selection speech synthesis and comprises a total of 1672 sentences. For the random selection of the 400 utterances to be recorded at both normal and fast speech rate phonological balance was not taken into account. The selected 400 sentences were recorded in two conditions:

- normal speech rate (approx. four syllables per second)
- fast and clear speech rate (approx. eight syllables per second) [1]

All recordings were conducted in a sound treated recording studio. Due to the fact that not all takes could be done in one session a strict monitoring of speaker and microphone position as well as of speaking rate, phrasing, accentuation, speaking style and intensity was necessary. As a consequence, speaker and microphone position were documented to ensure easy restoration. Additionally, several reference sentences were presented to the speaker in order to readjust her performance prior to each session as well as within the sessions. The reference sentences were utterances from the very first recording session reflecting the required speaking rate and style. All sentences were recorded at normal speaking rate first; only afterwards fast versions were elicited. To approach the fastest speaking rate possible, the speaker generally followed the strategy of repeating accelerated renditions of a sentence several times in a row. This procedure was shown to be useful by [Greisbach, 1992], [Liu and Zeng, 2006], and [Jannedy et al., 2010] who also trained their speakers to produce the required speaking style by gradually guiding them to the designated tempo. Thus, fast versions of one sentence were recorded repeatedly in succession with accelerated tempo and enhanced articulatory effort each time until the optimal combination of tempo and articulatory precision was reached. To record utterances at normal speech rate took approximately eight hours, recordings at fast speech rate, however, took three times longer.

Two phonetically skilled people supervised the recordings, gave instructions and feedback, and corrected the speaker immediately if necessary. After listening to all recorded fast versions of a certain utterance again, the realization perceived as being produced at the fastest speaking rate and at the same time articulated most clearly was selected by the supervisors to be included in the fast speech corpus. This way, two unit selection corpora were created: One at normal speech rate and one at fast speech rate articulated as accurately as possible.

7.1.1 Perceptual Evaluation

[Janse, 2003b] reported that artificially produced fast words whose temporal patterns were equivalent to natural fast speech were judged less intelligible than artificially produced fast words which were linearly compressed. The less

[1] Hereinafter referred to as "fast" speech, presuming that fast speech was articulated as accurately as possible during recordings.

a stimulus deviated from the canonical form, the better the word was understood by listeners. Taking these findings into account, the first goal of the current study was to perceptually evaluate whether the normal rate corpus recordings would indeed have an advantage regarding intelligibility and a disadvantage regarding naturalness compared to the fast rate corpus recordings when accelerated to the same fast speech rate. A first perception experiment was set up to evaluate stimuli featuring a speaking rate of approximately eight syllables per second ("fast" condition, [Moers et al., 2010c], [Moers et al., 2010d]). It was assumed that stimuli based on normal rate utterances would be perceived as more intelligible, but less natural than the unmodified fast sentences.

The second step was to find out whether fast speech recordings would have an advantage or disadvantage regarding intelligibility and naturalness compared to normal speech rate utterances in an "ultra-fast" condition as defined by [Moos and Trouvain, 2007]. To meet the criteria for this speaking rate condition, both normal and fast rate utterances had to be accelerated to an even faster and therefore highly unnatural speech tempo of approximately sixteen syllables per second [Moers et al., 2010c], [Moers et al., 2010d]. Hence, an overall decrease in naturalness judgments from the fast speaking rate condition evaluated previously to the ultra-fast condition investigated here was anticipated. However, sentences generated from normal rate recordings had to be modified more strongly with respect to their duration, whereas sentences generated from the fast speech rate corpus required a comparatively smaller duration manipulation. Therefore, in the ultra-fast condition stimuli generated from fast speech utterances were expected to be perceived as at least as intelligible as stimuli generated from normal rate utterances, but at the same time as more natural.

Thus, the first experiment was immediately followed by a second one to evaluate the intelligibility and naturalness of stimuli generated to match the "ultra-fast" speech rate condition [Moers et al., 2010c], [Moers et al., 2010d]. As the unnatural ultra-fast speaking rate of the stimuli presented in the second part of the experiment might draw the subjects' attention to intelligibility only - making it hard to judge naturalness independently - a third part was implemented where subjects were asked to only indicate the more natural sounding utterance of each ultra-fast stimulus pair again, but not to assign any naturalness score. It was anticipated that utterances generated from the fast speech rate recordings would be preferred over utterances based on the normal rate recordings. For stimuli of the first part of the experiment no such pairwise comparison of the perceived naturalness was conducted as it was assumed that the unmanipulated fast speech stimuli sounded more natural anyway.

Methods

For the first part of the experiment, twenty sentences were randomly picked from both corpora. The linguistic content of the picked normal and fast speech utterances was identical, though. Texts are listed in appendix C. In order to harmonize speaking rates as much as possible, the total duration of the normal and the corresponding fast rate utterance were measured in seconds, starting from the onset of the first sound to the offset of the last sound of the utterance. Afterwards, the ratio of the measured durations was calculated. This led to the definition of a durational factor describing the proportion between the normal and the fast realization of the respective sentence. Measured durations and durational factors for each pair of utterances are listed in table 7.1. The calculated durational factor was then applied to the normal rate utterances by means of the TD-PSOLA implementation in Praat [Boersma and Weenink, 2010]. This way, normal rate sentences were sped up linearly until they met the higher speech rate of the corresponding natural fast sentences [Moers et al., 2010c], [Moers et al., 2010d]. Although PSOLA is an algorithm which applies an unnatural linear manipulation to the signal, it was the method of choice here since [Janse, 2003a] noted that making the temporal pattern of artificially time-compressed speech (words) more similar to that of natural fast speech did not improve intelligibility compared to linear compression.

As can be derived from table 7.1, the average durational factor for most utterances was slightly higher than it was aimed for during corpus recordings: The set target was a proportion of approximately 0.5 between normal and fast rate utterances (cf. section 7.1), but the calculated mean for the evaluated samples was approximately 0.6 indicating that fast speech was not exactly twice as fast as the normal rate speech, but a little slower. However, this observation was not regarded having major implications for the results of the evaluation, since speaking rate variations across naturally produced sentences are a phenomenon which can be observed across all speaking rate conditions [van Santen, 1992], [Pfitzinger, 1998], [Wang et al., 2000]. Therefore, slight deviations in durational patterns could even be advantageous with regard to the perceived naturalness of the generated utterances [Moers et al., 2010c], [Moers et al., 2010d].

For the second part of the experiment, the ultra-fast condition, stimuli were generated on the basis of the same utterances as before. The calculated durational factor for the acceleration of the normal speech rate sentences was doubled whereas for the acceleration of the fast speech utterances, the durational factor was set to 2.0. Thus, while the speaking rate of the stimuli for the first part of the experiment was approximately eight syllables per second, the speaking rate of the stimuli for the second part of the experiment was adjusted to approximately sixteen syllables per second. The same ultra-fast stimuli were then reused for the third part of the evaluation [Moers et al., 2010c], [Moers et al., 2010d].

The experiment was implemented using the Praat ExperimentMFC envi-

Sentence ID	Duration normal speech [s]	Duration fast speech [s]	Durational factor normal to fast	Durational factor normal to ultra-fast
016	4.39	2.27	0.52	0.26
029	5.89	3.40	0.58	0.29
057	3.71	2.43	0.66	0.33
064	6.90	4.56	0.66	0.33
100	5.56	3.74	0.67	0.34
129	4.76	2.82	0.59	0.30
164	2.95	1.47	0.50	0.25
172	5.87	3.26	0.56	0.28
210	4.55	2.48	0.54	0.27
219	5.85	3.70	0.63	0.32
235	3.54	2.13	0.60	0.30
242	5.87	3.87	0.66	0.33
273	5.35	3.32	0.62	0.31
303	5.99	4.13	0.69	0.34
312	5.88	3.50	0.59	0.30
327	3.95	1.90	0.48	0.24
348	4.36	2.14	0.49	0.25
366	3.73	2.25	0.60	0.30
384	3.17	2.37	0.75	0.37
394	5.19	3.45	0.66	0.33

Table 7.1: Sentence duration for normal and corresponding fast speech utterances in seconds; durational factor applied to generate stimuli for the fast and ultra-fast conditions.

ronment [Boersma and Weenink, 2010]. Altogether, subjects were presented with sixty stimuli, each of them consisting of a pair of the same utterance generated from the two different underlying versions. One replay of each stimulus pair was permitted. The evaluation was conducted in a quiet environment, and stimuli were presented via earphones. Eleven subjects took part in the experiment. In the first two parts, subjects were instructed to indicate for each played stimulus pair the utterance which was pronounced more clearly and therefore was more intelligible. Immediately afterwards, they were asked to provide a naturalness rating for the more intelligible version of the stimulus pair, reaching from 1=*poor* to 5=*excellent*. Subsequently, in the third part of the experiment subjects were asked to indicate the more natural sounding utterance of each ultra-fast stimulus pair again without assigning a concrete naturalness score [Moers et al., 2010c], [Moers et al., 2010d].

Results

The approach to analyzing the results was similar to the one chosen in the perceptual evaluation of the speaker's fast speech described in section 6.2.2: The version of the utterance which was judged more intelligible received one point. This way, an intelligibility score was gained for each presented version of an utterance. Intelligibility scores per underlying speech rate version are depicted in figure 7.1: The two columns on the left represent intelligibility scores for stimuli based on normal speech (light grey) and for stimuli based on fast speech (dark grey) in the "fast" condition. The two columns in the middle, however, reflect intelligibility scores for the "ultra-fast" condition (normal speech=light grey, fast speech=dark grey). As expected, in the "fast" condition stimuli generated from normal speech rate recordings were judged more intelligible than natural fast ones (χ^2=5.25, df=1, p<.05). However, this advantage disappeared in the "ultra-fast" condition. There even was a slight tendency to prefer the stimuli generated from natural fast speech, albeit not a significant one. These results confirm the initial hypotheses regarding the intelligibility of the two different underlying speaking styles when linearly accelerated to fast or ultra-fast speaking rate, respectively.

Looking at the naturalness scores assigned during the first and the second part of the experiment, the advantage of natural fast speech stimuli was highly significant in the "fast" rate condition, as was expected: Stimuli consisting of natural fast speech were rated significantly more natural than stimuli generated from normal speaking rate utterances (Wilcoxon rank sum test with continuity correction, W=3855.5, p<.0001). However, this significant difference between naturalness scores disappeared in the "ultra-fast" condition. Comparing all naturalness scores assigned to stimuli presented in the "ultra-fast" condition to those assigned to stimuli in the "fast" condition, it became apparent that indeed naturalness scores for the "ultra-fast" condition were significantly lower than the ones for the "fast" condition (Wilcoxon rank sum test with continuity correction, W=38394, p<.0001). Nevertheless, when analyzing the results of

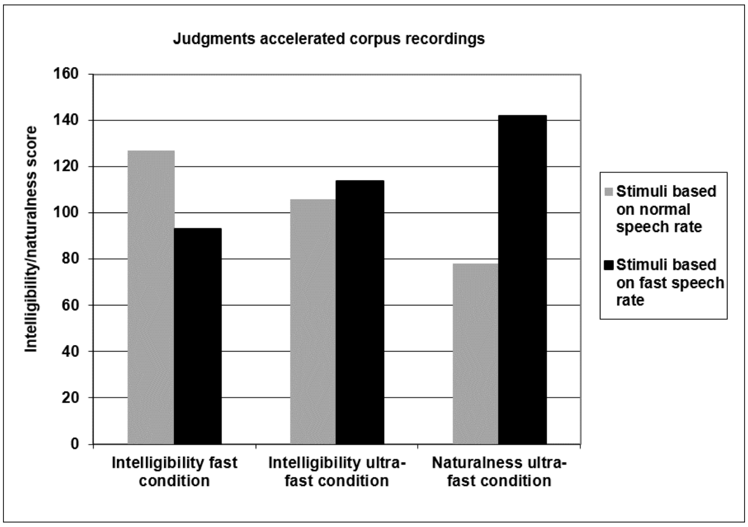


Figure 7.1: Corpus recordings: Intelligibility scores for fast and ultra-fast stimuli, naturalness scores for ultra-fast stimuli.

the third part of the experiment where subjects were asked to indicate the more natural sounding version of each ultra-fast stimulus pair again, it became evident that stimuli generated from natural fast speech were clearly preferred with respect to naturalness (χ^2=18.62, df=1, p<.0001). Figure 7.1 illustrates the results of the naturalness ratings from the third part of the experiment in its right most columns (normal speech=light grey, fast speech=dark grey). (cf. [Moers et al., 2010c], [Moers et al., 2010d]).

One important factor which may have influenced the outcome of this evaluation is the extensive manipulation of the normal rate versions which may have created artifacts known to appear when using the TD-PSOLA algorithm with a manipulating factor of two and more [Moulines and Charpentier, 1990], [Quené, 2007], [Liu et al., 2008], whereas stimuli based on clearly articulated fast speech needed less manipulation. At the same time, stimuli based on clearly articulated fast speech were assigned an intelligibility score comparable to the extensively manipulated normal speech rate utterances. Thus, stimuli based on fast speech had an advantage regarding naturalness and at least no disadvantage concerning intelligibility despite PSOLA manipulation [Moers et al., 2010c], [Moers et al., 2010d]. Another phenomenon probably influencing the results was observed by [Stent et al., 2011]: They stated that "listeners may not be very good at judging intelligibility" separated from other aspects. Thus, more detailed results could probably have been obtained here by evaluating the "Word Error Rate" as well. Taking all results of the intelligibility and naturalness judgments together, the findings confirm the initial hypotheses for the ultra-fast speaking rate condition, and therefore encourage the approach to use clearly articulated fast speech as a separate unit selection inventory for the synthesis of (ultra-)fast speech.

7.2 Inventory Preparation

The preparation of the unit selection inventory is a very time consuming step during the development of new corpora for unit selection synthesis systems. Therefore, automatic segmentation techniques are preferred over manual labeling. However, the quality of the synthesized speech largely depends on the accuracy of the label timing [Kominek et al., 2003], [Demenko et al., 2008]; cf. also [Chu et al., 2006]. If corpus recordings are based on fast speech, using the same segmentation algorithm for both normal and fast speech might result in a considerably increasing amount of incorrect labels for fast speech utterances. If this was the case, automatic phone segmentation would not be applicable to fast speech corpus recordings, even if the fast speech was articulated as accurately as possible. Consequently, the implementation of a fast speech inventory would not be desirable in terms of effort needed to prepare it for use in unit selection speech synthesis.

7.2.1 Automatic Segmentation

Methods

For automatic segmentation an HTK-based aligner adapted to German was used [Dragon, 2005]; cf. also [Young et al., 2006]. It was provided with the orthographic as well as with the canonically transcribed version (lexical form) of each recorded sentence. Prior to that, annotations in plain SAMPA [Wells, 1997] which were already available for all randomly selected "BITS corpus" sentences had to be adapted to the BOSS-SAMPA scheme [Breuer et al., 2001] to ensure applicability in the BOSS system afterwards [Moers et al., 2010c], [Moers et al., 2010a], [Moers et al., 2010d]. An average agreement of 94% between human labelers within a 20 ms tolerance interval for manual labeling of normal rate speech is reported in the literature [Adell et al., 2005], [Pfitzinger et al., 1996]; cf. also [Kawai and Toda, 2004]. This amount is regarded as a quality measure for automatic alignment techniques. Therefore, the window length for the alignment process was set to 20 ms as well. The processing lasted about two hours for each corpus version. In the end, there were 18.474 forced-aligned segments for normal and 18.231 forced-aligned segments for fast speech available.

To evaluate the accuracy of the automatic segmentation, the labeling of 49 randomly chosen sentences of each corpus was reworked manually using the sound visualization and manipulation tool WaveSurfer [Beskow and Sjölander, 2000], a public domain software which was able to read the label files produced by the HTK-based aligner. Manual label correction was done by only one person to maximize consistency. Each phone was listened to several times to minimize effects from neighboring phones. Boundaries between consonants and vowels were marked at the start of the modal voice at the beginning of the vowel and the fading of formants at the end of the vowel. This way, the boundaries of

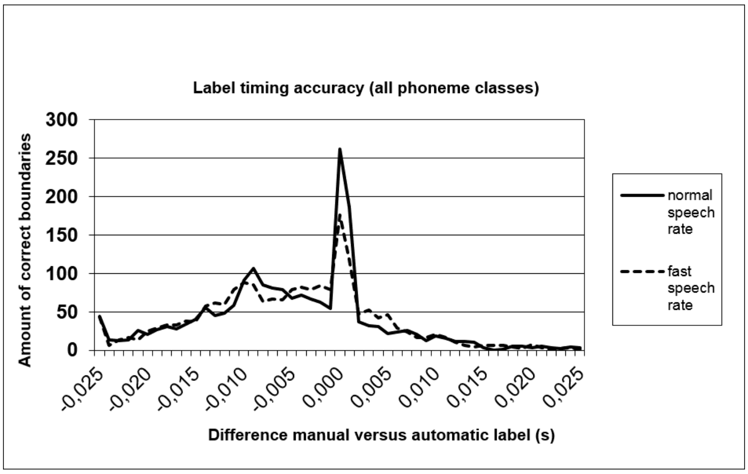


Figure 7.2: Frequency distribution of label timing accuracy for different intervals for normal (solid line) and fast (dotted line) rate speech.

2.091 segments of the normal rate corpus and 2.074 segments of the fast speech corpus were manually corrected. Afterwards, LTA was calculated for all phones in normal and fast rate utterances by subtracting the manual label time from the automatic label time. When the result was positive the automatically generated label was set too late with respect to the manual label; when the result was negative the automatic label was set too early. The frequency distribution of the LTA for different time intervals is plotted in figure 7.2 (normal speech = solid line, fast speech = dotted line) [Moers et al., 2010c], [Moers et al., 2010a], [Moers et al., 2010d].

Results

Results showed that for normal speech, 90.44% of the labels were set within the 20 ms tolerance interval. For fast speech, it was 90.80%. Although these results did not quite reach the quality criteria defined by [Adell et al., 2005], it was concluded that the approach of automatic segmentation in general was also applicable to fast speech, as the outcome for both speaking rate variants was almost identical. Moreover, [Demenko et al., 2010] revealed that speech generated from a fully-automatic segmented corpus was not perceived significantly worse than speech generated from a semi-automatically or a manually segmented corpus. Here, semi-automatic alignment included an amount of 30% manually inserted boundaries which doubled the objective labeling accuracy while the workload which was saved compared to full manual segmentation was still considerable. However, the semi-automatically segmented corpus was perceptually slightly inferior to both other corpora under investigation, although the error for automatic alignment was higher than the error for manual alignment. A suggested reason for this was that boundaries were most probably determined more consistently when doing automatic alignment.

Version	5 ms	10 ms	15 ms	20 ms
Normal	42.75%	68.82%	83.02%	90.44%
Fast	42.67%	65.67%	81.97%	90.70%

Table 7.2: Label timing accuracy in different tolerance intervals compared for a sub-corpus of 49 sentences of normal and fast speech.

Although the authors stated that more experiments were needed to decide why the semi-automatic approach delivered poorer results, it was derived from the insignificance between fully-automatic segmentation and manual segmentation that there was not much need to perform manual segmentation in any case.

However, since segment durations are known to be shorter in fast speech, for the work presented here it was decided to additionally analyze differences in LTA between the two speaking rate versions within several smaller time intervals. Moreover, it was evaluated whether decreased segment duration in fast speech had a negative influence on accuracy of labels. Results showed that this was not the case (cf. table 7.2). Thus, differences in LTA between normal and fast rate sentences were marginal across all tolerance intervals and not significant at all (Welch's t-test, p>.10), implying that automatic forced-alignment methods can be used for automatic segmentation of fast speech utterances as well. Nevertheless, the absolute count of correctly segmented labels at exactly 0 ms was significantly higher for normal (12.5%) than for fast rate speech (8.5%) (χ^2=16.886, df=1, p«.001) which can be attributed to increased coarticulatory effects and acoustic reduction occurring in fast speech making it more difficult to define the exact label boundary.

Since segment duration in fast speech is shorter overall, one might ask if the tolerance interval usually chosen to judge inter-labeler consistency for normal rate speech is appropriate to also judge the accuracy of fast speech label timing in general. Also an adaptation of the window length of the HTK-based aligner to fast speech segment durations should probably be taken into consideration for future work. However, since the overall results regarding correctly set boundaries within the commonly used 20 ms tolerance interval were very good for both normal and fast speech rate recordings, and moreover were also in accordance with the average agreement reported for human labelers, it was concluded that automatic phone segmentation is a technique applicable to recordings at both normal and fast speech rate, at least if the latter was performed with high precision and enhanced articulatory effort.

7.2.2 Duration Prediction

The duration of phonetic segments is an important prosodic factor in the production of natural sounding synthetic speech as well [Carlson et al., 1979], [Breen, 1992], [Brinckmann and Trouvain, 2003]. Considering the results of [Janse, 2003b] and [Brinckmann and Trouvain, 2003] as well as the outcome of the evaluation presented in section 7.1.1, it was decided to create a segmental duration prediction model for the normal as well as the fast speech corpus, repectively, by applying "CART" [Breiman et al., 1984], [Riley, 1990] ([Moers et al., 2010c], [Moers et al., 2010a], [Moers et al., 2010d]; cf. also section 3.2.1). Important phonetic and prosodic features influencing segmental duration were defined as features. It was hypothesized that the generated duration prediction model would show higher correlations between observed and predicted durations with normal speech rate data than with fast speech rate data because of the higher degree of coarticulation and reduction not completely avoidable when producing fast speech, even if it was articulated as accurately as possible. If the differences in correlations were significant, a CART-based duration prediction model would not be applicable for the fast speech inventory.

Methods

The tool applied to generate CART-based duration prediction models for the normal and the fast speech corpus was *wagon* from the Edinburgh Speech Tools [King et al., 2003]. The feature set applied was adapted to the requirements of the unit selection synthesis system BOSS [Breuer et al., 2005], taking into account not only the most important features influencing segmental duration, but also other prosodic factors [Breuer et al., 2006b]:

- phone identity
- phone duration
- preceding phone
- next phone
- one after next phone
- phrase position
- syllabic stress

The phoneme itself was the feature whose duration was to be predicted. The phone durations extracted from the particular corpus were the training data. The position in the phrase was either "initial", "medial" or "final". Syllabic stress had one of the values "primary", "secondary" or "none". The phoneme itself, its duration, the preceding and following phoneme as well as the syllabic stress were extracted from the corpus after it was pre-processed by the applicable "BOSS-Tools" [Breuer et al., 2005], [Breuer et al., 2006b]. The second following phoneme and the phrase position had to be calculated during further processing [Moers et al., 2010c], [Moers et al., 2010a], [Moers et al., 2010d].

Results

The performance of a duration prediction model is usually measured by comparing predicted segmental durations to durations observed in the database. An alternative way of evaluation would be a perception experiment where

Version	Correlation	RMSE	Mean (abs) error
Normal	0.80	39.66	20.16 (34.16)
Fast	0.78	23.97	12.37 (20.53)

Table 7.3: CART duration prediction results for both corpora (RMSE and Mean (abs) error in ms).

Feature	Correlation normal speech rate corpus (dataset of 18487 vectors of 7 parameters)	Correlation fast speech rate corpus (dataset of 18240 vectors of 7 parameters)
phone identity	**0.4734**	0.7559
position in phrase	0.6750	0.6649
next phoneme	0.7862	**0.4736**
preceding phoneme	0.8000	0.7681
syllabic stress	0.8009	0.7738
next but one phoneme	0.8018	0.7749

Table 7.4: Feature ranking comparison (stepwise) of CART results for both corpora.

listeners judge the outcome of different duration models applied in synthetic speech synthesis [Brinckmann and Trouvain, 2003]. As it was not planned to generate and compare different duration prediction models for fast speech, the approach to analyzing the quality of the duration prediction model by comparing predicted segment durations to observed segmant durations was preferred here. Results showed that the correlation between observed and predicted duration for fast rate utterances was 0.78 whereas the correlation between observed and predicted duration for normal rate utterances was 0.80 (cf. table 7.3). This was only a slight difference; both correlations were similar to results reported for the prediction of segmental duration in normal speech rate in other languages [Riley, 1990], [Batusek, 2002], [Krishna and Murthy, 2004], [Chung and Huckvale, 2001], [Klessa et al., 2007], [Demenko et al., 2010]. Root Mean Square Error (RMSE) and Mean (absolute) error were even smaller for the fast speech corpus. However, this probably had to be attributed to the fact that overall segment durations were smaller for fast speech.

However, looking at the feature ranking generated by means of the *stepwise* option of ”wagon” [King et al., 2003] and listed in table 7.4, major differences between the two duration prediction models become apparent. The most important feature for the prediction of the duration of a phoneme in normal rate speech is the phoneme itself; for fast speech, it was the phoneme following the phoneme who’s duration was to be predicted. This might be due to stronger coarticulatory effects attributed to increased articulatory overlap in fast speech. Syllabic stress surprisingly shows only marginal differences for normal versus fast speech. Since the total number of stressed syllables and their duration generally decrease in fast speech, it was expected that syllabic stress would show a higher impact on the correlation between observed and predicted durations for normal rate speech than for fast rate speech [Moers et al., 2010c], [Moers et al., 2010a], [Moers et al., 2010d]. This was not the case, which might be attributed to the enhanced articulatory effort during fast speech production. However, the *wagon* manual states that the feature ranking generated by the *stepwise* option shall not be used to derive a general order of importance of the selected features.

Taking these findings together, the duration prediction models showed very similar correlations between observed and predicted durations for both normal and fast rate speech, although the feature space was not too manifold. Therefore, it was concluded that building a CART-based segmental duration prediction model is applicable to normal as well as fast speech corpora. A further refinement of the feature set, e.g. by including place and manner of articulation, as well as the inclusion of supra-segmental features like position in the foot or word and sentence stress may enhance the accomplished prediction accuracy in the future (cf. [Möbius and van Santen, 1996], [Brinckmann and Trouvain, 2003]).

7.3 Summary and Conclusions

Corpus recordings were performed with the selected speaker. Two parallel corpora were recorded: One at normal and one at fast speech rate articulated as accurately as possible. A perceptual evaluation of the corpus recordings confirmed observations already reported by [Janse, 2003b] for the ”fast” speech condition (eight syllables per second): Stimuli generated from normal speech rate recordings were judged more intelligible than natural fast ones. In the ”ultra-fast” condition (sixteen syllables per second), however, there was a slight tendency for listeners to prefer stimuli generated from fast speech recordings with respect to intelligibility, and a significant preference with respect to naturalness. Nonetheless, the stimuli based on normal rate speech may have suffered more strongly from the modification of the speech signal imposed by TD-PSOLA ([Moulines and Charpentier, 1990], [Liu et al., 2008]), which in turn may have influenced the naturalness judgments adversely. Still, it was decided to use PSOLA to manipulate speaking rate here, because it is still generally applied in speech synthesis systems. An alternative approach to be evaluated in future research would be the application of other acceleration algorithms, e.g. non-linear time scaling as proposed by [Höpfner, 2008].

The automatic phone segmentation conducted afterwards by means of an HTK-based aligner adapted to German [Dragon, 2005] showed only marginal differences in label timing accuracy for normal versus fast speech. Given the satisfactory segmentation performance within the commonly applied 20 ms tolerance interval for both speaking styles and significantly shorter segment durations in fast speech, it was concluded that automatic phone segmentation is a technique applicable to recordings at both normal and fast speech rate, at least if the latter was performed with high precision and enhanced articulatory effort. Nevertheless, an optimal strategy for improving alignment accuracy might be the adaptation of the window length for fast speech in the segmentation algorithm, since segment durations in fast speech are shorter in general. Thus, one might ask if the tolerance interval chosen here is appropriate to judge the accuracy of fast speech label timing.

CART-based duration prediction models generated for both corpora independently while considering important phonetic and prosodic features influencing segmental duration revealed that the correlation between observed and predicted segment duration was comparable for recordings at both speaking rates. It was concluded that this technique is applicable for normal as well as fast and clear speech utterances. However, as slight differences in correlations and feature ranking between normal and fast speech recordings were observed, the model may require a refinement of features applied to enhance the correlation used to predict duration patterns of fast speech. A larger database or more features might be a solution here, as suggested by [Klessa et al., 2007] and [Demenko et al., 2010]. Also a comparison of the canonical or lexical form versus the transcription of the actually realized form might reveal additional insights. Moreover, perception experiments comparing different prediction models might be more meaningful, since objective acoustic measures do not always reflect subjective perception [Brinckmann and Trouvain, 2003]. Taking all findings together, the idea to implement fast speech as a separate corpus in a unit selection synthesis system was further supported. Therefore, fast speech was actually implemented in the BOSS system.

Chapter 8

Synthesizing Ultra-Fast Speech

According to the results of the preliminary evaluation outlined in chapter 5, the blind and visually impaired who rely on speech output when using a computer or other technocal devices often prefer a speaking rate which goes far beyond what is naturally producible when speaking fast (cf. [Moos and Trouvain, 2007], [Adank and Janse, 2009]). Speech synthesis systems which are based on formant synthesis are able to generate such high speaking rates, but the generated speech sounds very unnatural, even when synthesized at a normal speaking rate. In contrast, unit selection speech synthesis systems usually produce output which is perceived as much more natural [Black and Taylor, 1997]. However, up to now the implementation of natural fast speech as unit selection inventory was not taken into consideration. Instead, speech synthesized with a unit selection synthesis system often is heavily manipulated through algorithms like TD-PSOLA, changing the duration of the speech segments to produce fast speech output. Therefore, speech generated with unit selection synthesis at ultra-fast speaking rates loses its advantage over formant synthesis regarding naturalness.

Nevertheless, trained blind and visually impaired people preferred the less natural sounding formant synthesis over the more natural sounding diphone synthesis across all speaking rates from ”normal” to ”ultra-fast” (cf. chapter 5, [Moos and Trouvain, 2007]). Aside from pure habituation due to repeated exposure (cf. [Jannedy et al., 2010], chapter 4.2), the unproblematic replication of fast and smooth transitions in formant synthesis as opposed to diphone concatenation may have played a vital role in this preference (cf. [Fowler, 2005]). However, as [Winters and Pisoni, 2004] pointed out, the advantage of formant synthesis might disappear when concatenative synthesis with larger units is used. Therefore, the next step after implementing fast and clear speech as an independent unit selection inventory in the BOSS system was to define the adequate unit size to synthesize fast speech in an optimal way. Since the acoustic transitions between consecutive segments are very important for the intelligibility of speech in general [Martinez et al., 1997], as well as for the intelligibility of synthetic speech in particular [Peterson et al., 1958], [Amerman and Parnell, 1981], [Janse, 2003a], discontinuities introduced to synthesized speech by concatenation should be minimized. As a consequence, [Breuer and Abresch, 2004] suggested to treat phone sequences which are prone to heavy coarticulation as atomic in the sense that they are regarded as two or more phones, but one indivisible synthesis unit which they called "phoxsy units" (phone extensions for synthesis). This approach was picked up in the investigation about defining the adequate unit size presented in section 8.1. It was expected to find a possible solution to modeling fast speech both more naturally (by using prerecorded concatenation units) and more intelligibly (by including typical smooth transitions in heavily coarticulated contexts) in order to achieve synthetic speech that was both maximally natural and maximally fast.

Afterwards, the intelligibility, naturalness, and overall acceptability of utterances generated from different underlying corpora and different systems at different speaking rates were evaluated. To investigate users' preferences, Semantically Unpredictable Sentences (SUS, [Benoit and Grice, 1996]) were generated by means of both unit selection inventories as well as formant synthesis (cf. [Syrdal et al., 2012]). Applied SUS are listed in appendix E. Afterwards, a 5 point Mean Opinion Score (MOS) was collected from two different listener groups: trained blind and visually impaired users of a specific screenreader software and untrained naive listeners (mostly sighted). Untrained listeners were included as a control group to accommodate the possible bias for the trained blind and visually impaired users regarding formant synthesis. Additionally, the Word Error Rate (WER) was analyzed depending on listener group and synthesis system. The results of the perceptual evaluation of speech synthesized from different underlying corpora and different systems at different speaking rates are outlined in detail in section 8.2. It was hypothesized that trained blind or visually impaired people would generally judge stimuli generated with formant synthesis better than stimuli generated from either unit selection corpus. Additionally, ultra-fast stimuli were anticipated to get a better MOS from blind listeners than from sighted listeners. Moreover, it was expected that for the trained listener group the WER for ultra-fast stimuli was significantly lower for formant synthesis based stimuli than for unit selection based stimuli. Regarding the sighted, untrained control group it was expected that stimuli generated from the fast speech unit selection corpus would get a higher MOS than stimuli generated from the normal speech unit selection corpus as well as stimuli generated with formant synthesis as the intelligibility of the fast speech unit selection based stimuli would be comparable to the intelligibility of other stimuli groups, but their naturalness would be higher.

IPA	BOSS-SAMPA	phoxsy unit
[?] + vowel	? + vowel	? + vowel
[h] / [ɦ] + vowel	single phones	h + vowel
[j] + vowel	single phones	j + vowel
[ʊ] / [v] + vowel	single phones	v + vowel
[ʀ] / [ʁ] / [ɾ] / [r] + vowel	single phones	r + vowel
[l] + vowel	single phones	l + vowel
[ən] / [n]	@n	@n
[əm] / [m]	single phones	@m
[əl] / [l]	single phones	@l
[j] / [ʊ] / [v] / [ʀ] / [ʁ] / [ɾ] / [r] / [l] + [ən]	single phones	j / v / r / l + @n
[j] / [ʊ] / [v] / [ʀ] / [ʁ] / [ɾ] / [r] / [l] + [əm]	single phones	j / v / r / l + @m
[j] / [ʊ] / [v] / [ʀ] / [ʁ] / [ɾ] / [r] / [l] + [əl]	single phones	j / v / r / l + @l
[ts]	ts	ts
[pf]	pf	pf

Table 8.1: Unit definitions in IPA, BOSS-SAMPA and as phoxsy units (after [Breuer and Abresch, 2004]).

8.1 Finding the Adequate Unit Size

8.1.1 Phoxsy Units

In the field of unit selection synthesis it is well known that linguistically motivated units like phones do not provide optimal properties for concatenation. The main disadvantage of this type of unit is the disregard of acoustic and auditive continuity. Phone extensions for synthesis ("phoxsy units") are therefore defined to systematically avoid concatenation points in the signal at positions where they are highly undesirable [Breuer and Abresch, 2004], [Breuer, 2009]. Essentially, they are sequences of phones prone to heavy coarticulation with fluent transitions and phonetically non-existing boundaries.

For the current investigation, phoxsy units were implemented as an independent multi-phone unit level in the BOSS ([Klabbers et al., 2001], cf. also section 3.1.2) in order to provide a robust and accessible usage [Moers et al., 2010b]. The modular architecture of BOSS allowed an unproblematic integration of the new multi-phone level into the existing system. The BOSS tool blf2xml, which extracts information from the BOSS Label Format files "blf" [Breuer et al., 2001] and creates an XML database, has been extended in order to recognize phoxsy units using the BOSS-FSA class (a finite state automaton). The tool additionally inserted the required units into the XML database. Other BOSS tools have also been adapted to calculate additional unit information like context classes, phrasing information, and MFCCs for phoxsy units to add them to the XML database. By means of the blfxml2db tool, the new multi-phone level was inserted into a MySQL database while calculating the unit index. The unit index is a unique number which identifies every unit in the corpus. Mapping tables then provide links between the units of two adjacent levels. Those levels are arranged hierarchically from words over syllables to phones and half-phones. The phoxsy multi-phone level was implemented as an intermediate level between syllables and phones. To maintain the hierarchy of the unit levels, a complete coverage of the corpus by phoxsy units was necessary. In addition, also the syllable map had to be adapted n order to provide links between syllables and phoxsy units instead of syllables and phones. Moreover, a phoxsy unit map had to be generated in order to provide links between phoxsy units and phones. A new preselection file for multi-phone unit preselection had to be created accordingly. The BOSS-Unitselection class was adapted and a new level PHOXSY was added to the BOSS-Node class. Also the BOSS-Transcription class was adapted to identify and insert phoxsy units into the internal system communication structure. It used the same mechanism as the blf2xml tool. For further details on BOSS modules and their interaction refer to [Breuer and Hess, 2010].

Table 8.1 lists possible phone combinations defined as phoxsy units by [Breuer and Abresch, 2004]. The ”IPA” column shows the unit definitions transcribed according to the International Phonetic Alphabet [International Phonetic Association, 2005]. The ”BOSS-SAMPA” column shows the way how the units have been processed in BOSS before phoxsy units were defined, whereas the ”phoxsy” column shows the new unit definitions in BOSS-SAMPA notation, which is a modified X-SAMPA notation [Breuer, 2009]. It was expected that not only speech synthesized from the normal rate unit selection corpus would benefit from the use of phoxsy units, but that particularly the intelligibility of utterances synthesized from the fast and clear speech corpus would be much higher with than without the use of phoxsy units.

8.1.2 Perceptual Evaluation of Speech Synthesized at Normal Speech Tempo

Methods

As a first step, the advantage of using phoxsy units as an additional unit level for normal rate speech synthesis as presented by [Breuer and Abresch, 2004] had to be verified. Therefore, after corpus preparation and implementation (cf. chapters 7.2 and 8.1.1), fifteen sentences from different possible application domains were synthesized. The text of the fifteen sentences and the respective domains are documented in appendix D. Each of the sentences contained at least three phoxsy units (marked in bold in appendix D). All utterances were synthesized using the normal speech rate corpus by applying four different strategies (cf. [Moers et al., 2010b]):

- Use of phones only
- Use of phoxsy units only
- Use of all unit levels excluding phoxsy units
- Use of all unit levels including phoxsy units

This way, sixty stimuli were generated to be evaluated by listeners. As a pairwise comparison between all four versions of a synthesized sentence would have exceeded a reasonable amount of listening tasks, it was decided to split the test sentences into two subsets, one comparing stimuli generated from a single unit level (phones only versus phoxsy units only), and another subset comparing stimuli generated from all unit levels excluding or including phoxsy units.

Thus, the first experiment was a pairwise comparison between stimuli synthesized by using only phones and stimuli synthesized by using only phoxsy units. Fourteen subjects took part in the experiment. All of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version of the played sentence was more intelligible and which version sounded more natural. Since [Breuer and Abresch, 2004] reported a high level of inconsistency in ratings which they ascribed to the great acoustic similarity of the stimuli they used for their experiment, each of the fifteen sentences was presented twice to assess reliability of judgments. The second evaluation was a pairwise comparison between stimuli synthesized from all unit levels excluding phoxsy units and stimuli synthesized from all unit selection levels including phoxsy units. Twenty-four subjects took part in this experiment. All of them were naive listeners and not experienced in using speech synthesis. Again, subjects were instructed to indicate for each stimulus pair the utterance which was more intelligible. Immediately afterwards, they were asked to judge which version sounded more natural. To evaluate consistency of ratings, also here each of the fifteen stimulus pairs was presented twice to the listeners.

Both experiments were implemented using the Praat ExperimentMFC environment [Boersma and Weenink, 2010]. Altogether, subjects were presented with thirty stimuli per experiment, each of them consisting of a pair of the same utterance generated by means of two different competing unit selection strategies using different unit levels. One replay of each stimulus pair was permitted. The experiment was conducted in a quiet environment and stimuli were presented via earphones. It was hypothesized that utterances containing either phoxsy units only or all unit levels including phoxsy units would be judged more intelligible than utterances generated from phones only or all unit selection levels excluding phoxsy units. Also naturalness was expected to benefit from the use of phoxsy units, since less concatenation points would be necessary. The approach to analyzing the results was similar to the one chosen for the perceptual evaluation of the speaker's fast speech presented in section

6.2.2: The version of the sentence which was judged more intelligible received one point. This way, an intelligibility score was gained for each unit selection version underlying the presented stimuli. The same method was applied to naturalness ratings (cf. [Moers et al., 2010b]).

Results

Figure 8.1 (top) shows the "more intelligible" and "more natural" judgments for stimuli consisting of phones only (dark grey columns) versus stimuli consisting of phoxsy units only (light grey columns), based on the normal rate speech corpus. For reasons of comparability, scores are plotted as average per subject. Results showed a significant difference (χ^2=15.66, df=1, p«.0001) for intelligibility judgments where stimuli using phoxsy units were rated as more intelligible than stimuli based on single phones. For naturalness judgments, no significant difference between the two versions was found.

Taking a closer look at the reliability of judgments it appeared that only thirteen of fourteen subjects rated intelligibility consistently above chance level. The one exception showed a rating consistency below 60% which was interpreted to indicate random preference. For naturalness judgments, the discrepancy was even higher: Only seven subjects rated naturalness consistently above chance level. As a consequence, inconsistent ratings were removed from the results. The outcome of this approach is depicted in figure 8.1 (bottom). It again shows the average "more intelligible" and "more natural" scores for stimuli consisting of phones only (dark grey columns) versus stimuli consisting of phoxsy units only (light grey columns), based on the normal rate speech corpus. Also here, results showed a significant difference (χ^2=18.57, df=1, p«.0001) for intelligibility judgments, but no significant difference regarding naturalness.

Results of the second experiment are depicted in figure 8.2 (top). It again shows the average "more intelligible" and "more natural" score of speech generated at normal speech rate. Ratings of stimuli synthesized using all unit levels excluding phoxsy units (dark grey columns) are plotted against ratings of stimuli synthesized including phoxsy units (light grey columns). Also here, the analysis revealed a significant difference (χ^2=9.70, df=1, p<.01) for intelligibility judgments where stimuli using phoxsy units were rated as more intelligible than stimuli synthesized excluding phoxsy units. For naturalness judgments, again no significant difference between the two versions was found.

In this scenario, the number of reliable intelligibility and naturalness judgments decreased dramatically when rating consistency was taken into account. Only eight out of twenty-four subjects revealed a rating consistency above 60% for intelligibility of normal rate stimuli. For naturalness judgments, the decrease was less dramatic: Twelve subjects rated naturalness consistently above chance level. The outcome of removing the inconsistent results is depicted in figure 8.2 (bottom). It again shows the average "more intelligible" and "more natural" scores for stimuli consisting of phones only (dark grey columns) versus

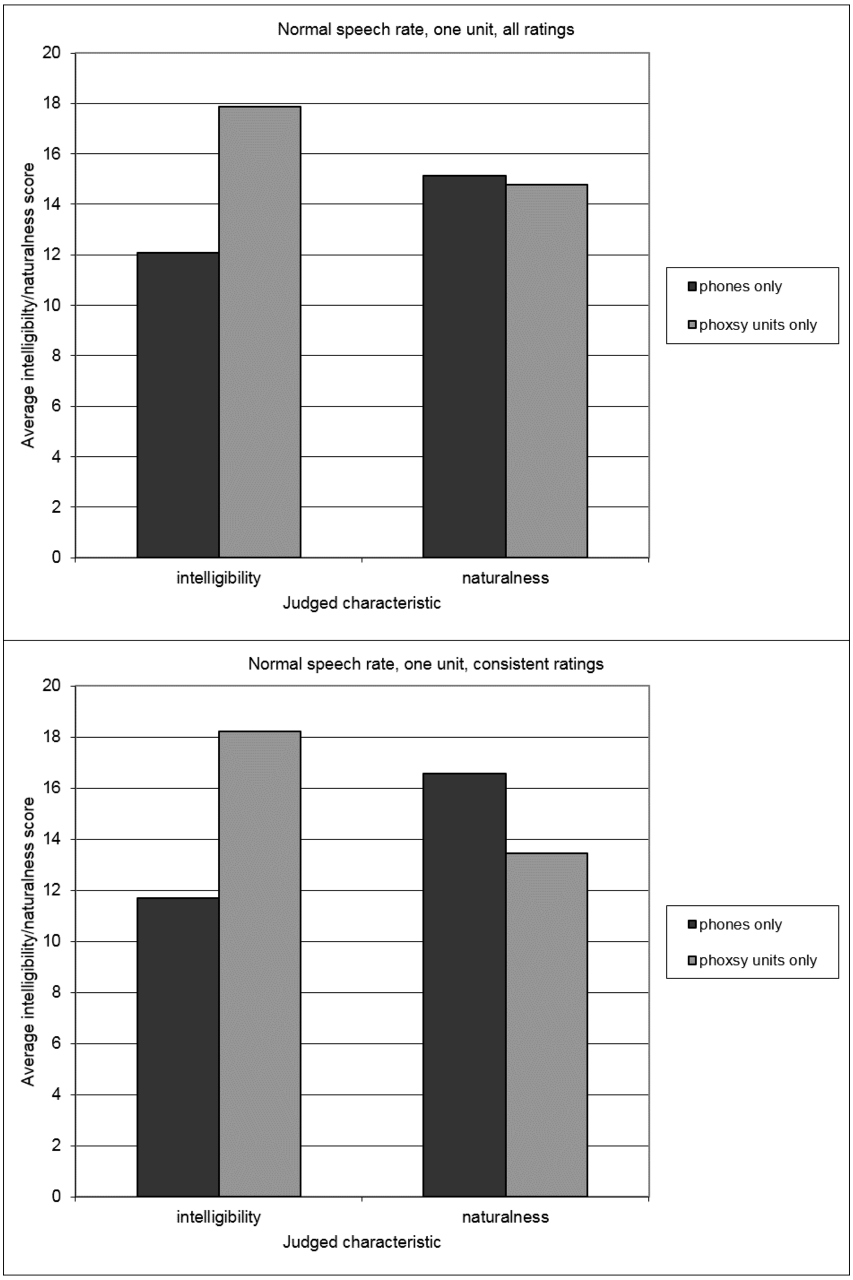


Figure 8.1: Average intelligibility and naturalness scores for normal speech stimuli generated using either phones only (dark grey columns) or phoxsy units only (light grey columns). Top: all ratings, bottom: consistent ratings.

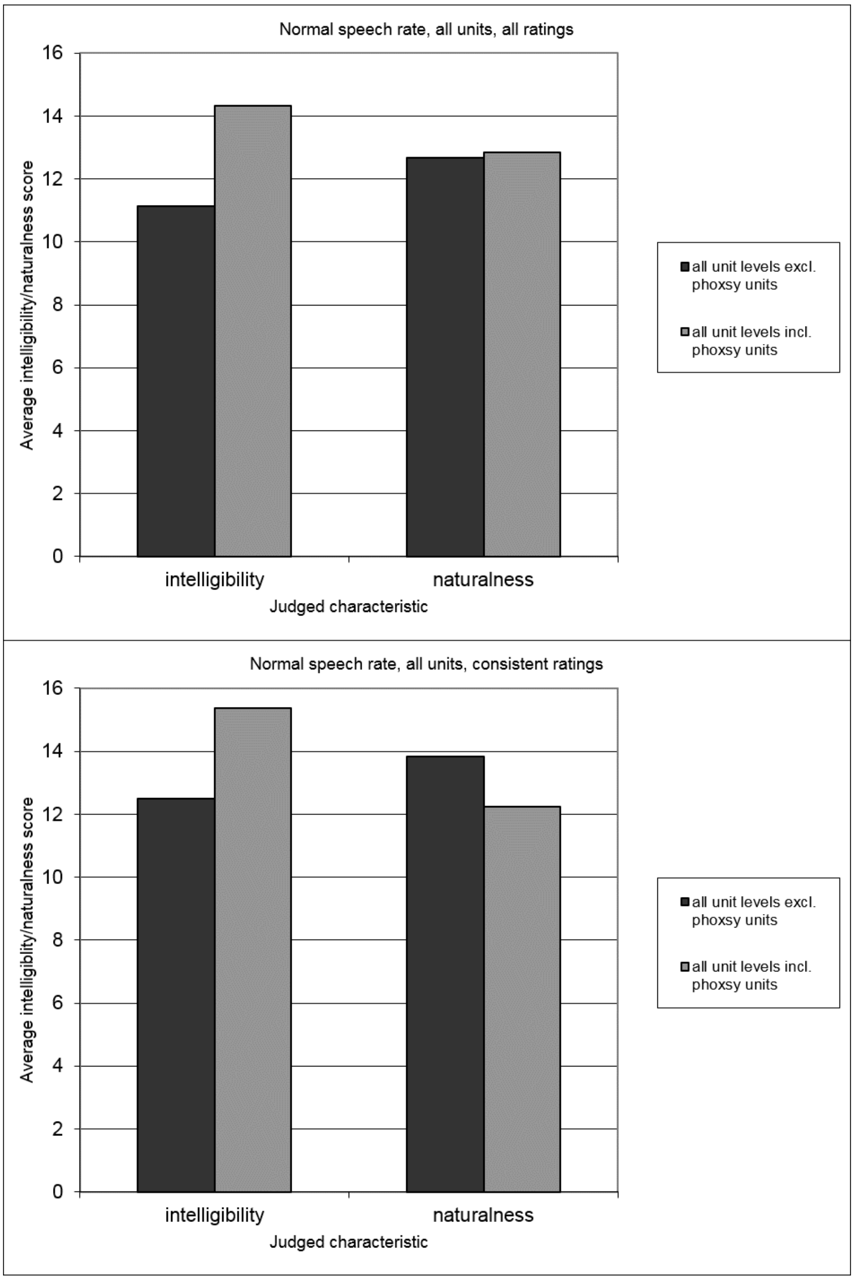


Figure 8.2: Average intelligibility and naturalness scores for normal speech stimuli generated using either all units levels excluding phoxsy units (dark grey columns) or including phoxsy units (light grey columns). Top: all ratings, bottom: consistent ratings.

stimuli consisting of phoxsy units only (light grey columns), based on the normal rate speech corpus. Due to the high inconsistency in judgments, it was not completely unexpected that the results did not show a significant difference anymore between the different synthesis strategies, neither for intelligibility nor regarding naturalness (cf. [Moers et al., 2010b]).

To sum up, in partial agreement with the current expectations the evaluation of synthesized normal rate speech indeed showed a significant advantage in intelligibility for stimuli generated from phoxsy units only and stimuli generated from all unit levels including phoxsy units compared to stimuli where phoxsy units were not considered for synthesis. However, naturalness judgments did not reveal a significant difference between unit selection approaches. Leaving out judgments which were inconsistent and near chance level, the advantage of phoxsy units with regard to intelligibility disappeared. Naturalness judgments were even slightly lower for utterances synthesized by means of phoxsy units, but still not to a significant amount. Thus, the rather ambiguous results regarding the use of phoxsy units to synthesize normal rate speech presented by [Breuer and Abresch, 2004] were substantiated for utterances generated from the normal rate unit selection corpus implemented here.

8.1.3 Perceptual Evaluation of Speech Synthesized at Fast Speech Tempo

Methods

Again, fifteen sentences containing at least three phoxsy units were synthesized, this time on the basis of the fast and clear speech corpus. To generate stimuli, the same four strategies as for the synthesis of the normal rate speech stimuli of the first experiments were applied:

- Use of phones only
- Use of phoxsy units only
- Use of all unit levels excluding phoxsy units
- Use of all unit levels including phoxsy units

The text of the fifteen sentences and the respective domains were identical with texts and domains used for the evaluation of the normal rate speech before (cf. appendix D). Thus, another sixty stimuli were to be judged by listeners. It was expected that utterances containing either phoxsy units only or all unit levels including phoxsy units would be perceived as more intelligible than utterances generated from phones only or all unit levels excluding phoxsy units, because phoxsy units provide more contextual information than single phones and would therefore cover even better for coarticulation and reduction phenomena occurring in fast speech. As less concatenation points would be necessary when using phoxsy units, it was hypothesized that naturalness would benefit from the use of phoxsy units as well. However, since phoxsy units are defined as sequences of phones prone to heavy coarticulation which is not completely avoidable during the production of fast speech - even if it is produced with high precision and enhanced articulatory effort - their use may as well have a considerable negative impact on the intelligibility and naturalness of fast speech synthesized from a fast speech unit selection inventory. Thus, a possible yet undesirable effect of using phoxsy units may be a degrading intelligibility of the generated speech due to comprised natural coarticulation and reduction phenomena instead of enhancing the intelligibility and/or naturalness of synthesized fast speech.

Again, the first experiment comprised a pairwise comparison of stimuli generated from the fast speech inventory by using only phones for synthesis and stimuli generated by using only phoxsy units. Twenty-two subjects took part in the experiment. All of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version of each stimulus pair presented was more intelligible and which one sounded more natural. Each of the fifteen sentences was presented twice to check for consistency of ratings. In accordance with the evaluation of the normal rate speech, the second experiment of synthesized fast speech evaluation was a pairwise comparison of stimuli synthesized from the fast speech inventory by using all unit levels excluding phoxsy units and stimuli synthesized by using all unit levels including phoxsy units. Fourteen subjects took part in the experiment. All of them were naive listeners and not experienced in using speech synthesis. Subjects were asked to indicate which version of the sentence was more intelligible and which version sounded more natural. Each of the fifteen sentences was presented twice to also here assess reliability of judgments afterwards [Moers et al., 2010b].

Results

Figure 8.3 shows the ”more intelligible” and ”more natural” judgments for fast speech stimuli consisting of phones only (dark grey columns) as opposed to stimuli consisting of phoxsy units only (light grey columns). In contrast to speech generated at normal speech tempo, results here showed a significant difference (χ^2=312.39, df=1, p«.0001) for intelligibility judgments as well as for naturalness judgments (χ^2=64.89, df=1, p«.0001).

The analysis of the reliability of ratings revealed consistent judgments for all twenty-two subjects regarding intelligibility. For naturalness judgments, however, the number of inconsistencies was much higher: Only seventeen listeners rated naturalness consistently above chance level. As a consequence, inconsistent ratings were removed from the results. Figure 8.3 (bottom) displays the results of this approach. The average ”more intelligible” and ”more natural” scores for stimuli consisting of phones only (dark grey columns) versus stimuli consisting of phoxsy units only (light grey columns) are depicted. As there were no inconsistent ratings for intelligibility, the statistic analysis revealed the same significant difference (χ^2=312.39, df=1, p«.0001) as before.

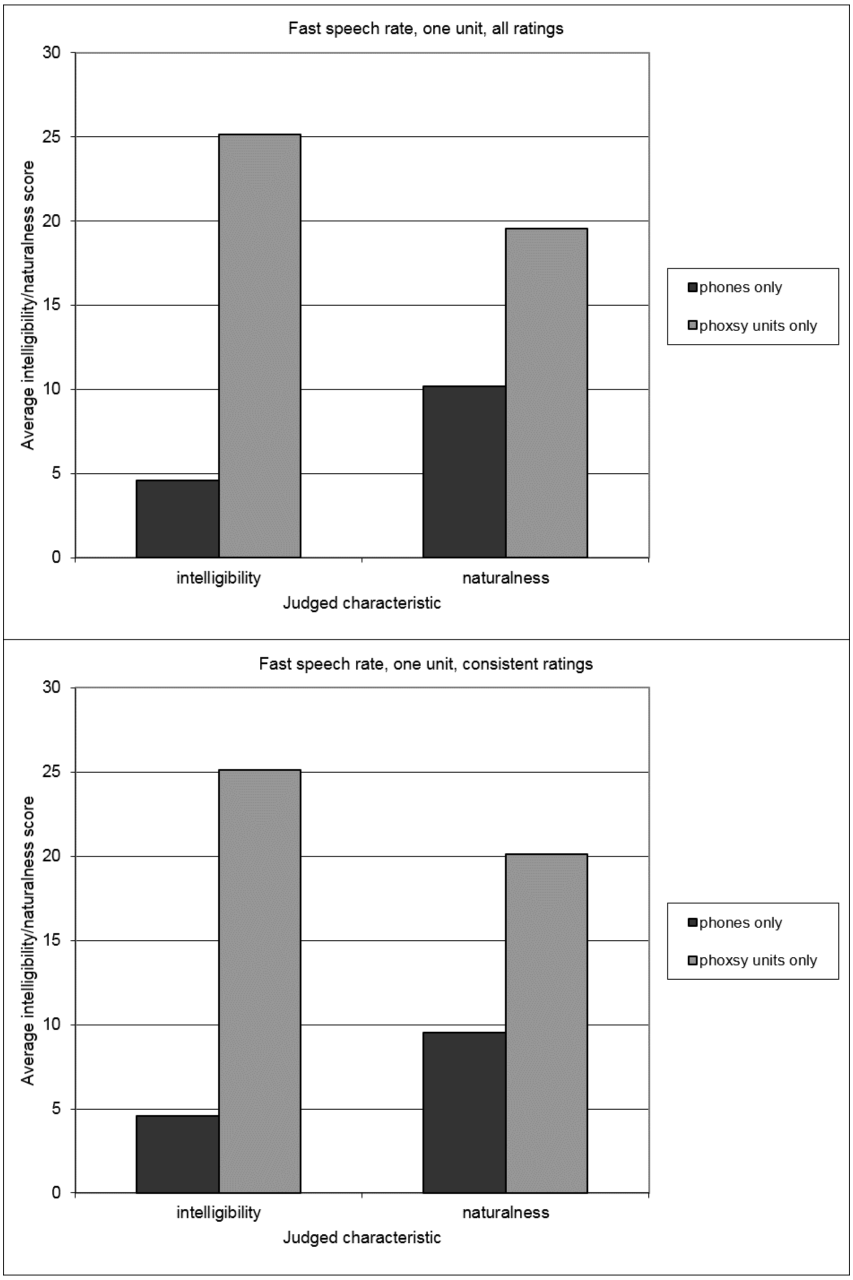


Figure 8.3: Average intelligibility and naturalness scores for fast speech stimuli generated using either phones only (dark grey columns) or phoxsy units only (light grey columns). Top: all ratings, bottom: consistent ratings.

Also the highly significant difference between the two versions regarding naturalness persisted (χ^2=64.29, df=1, p«.0001).

Figure 8.4 shows the ”more intelligible” and ”more natural” scores for fast stimuli generated from all unit levels excluding phoxsy units (dark grey columns) versus stimuli consisting of all levels including phoxsy units (light grey columns). Again, results revealed a significant difference (χ^2=24.89, df=1, p«.0001) for intelligibility judgments. For naturalness judgments, no significant difference between the two versions was observed; ratings were almost at chance level. Surprisingly, thirteen of fourteen subjects rated intelligibility as well as naturalness of fast speech stimuli generated from all unit levels, either excluding or including phoxsy units, consistently above chance level. Resulting intelligibility and naturalness scores are depicted in figure 8.4 (bottom). Stimuli consisting of single phones are depicted in dark grey columns, stimuli consisting of phoxsy units only in light grey columns. Also here, results showed a significant difference (χ^2=22.35, df=1, p«.0001) for intelligibility judgments, but no significant difference between the two versions regarding naturalness, despite a slight tendency to prefer stimuli generated from all unit levels including phoxsy units (cf. [Moers et al., 2010b]).

The evaluation of fast speech synthesized from a fast and clear speech inventory showed a significant advantage in both intelligibility and naturalness for stimuli generated by means of phoxsy units only. For stimuli generated from all unit levels excluding or including phoxsy units, respectively, a significant difference was found for intelligibility judgments as well. Since the analysis of judgments of synthesized fast speech revealed more significant results than for normal rate speech, it was concluded that multi-phone (”phoxsy”) units are not only applicable to slightly enhance the intelligibility of speech synthesized from a normal speech rate inventory, but to improve even more the intelligibility and to some extent the naturalness of fast speech synthesized from an independent fast and clear speech inventory.

8.2 Intelligibility and Naturalness of Synthesized Ultra-Fast Speech

After finding the adequate unit size to synthesize fast speech and confirming previous findings by [Breuer and Abresch, 2004], a perceptual evaluation of speech generated at different speaking rates based on the fast and clear speech unit selection corpus compared to speech generated at different speaking rates based on the normal rate speech unit selection corpus as well as to speech synthesized at different speaking rates with the formant synthesis system ”JAWS Eloquence” [FreedomScientific, 2011] was conducted (cf. [Moers, 2011]). Synthesized utterances were evaluated with regard to intelligibility, naturalness, and overall acceptability. Although [Jongenburger and van Bezooijen, 1992],

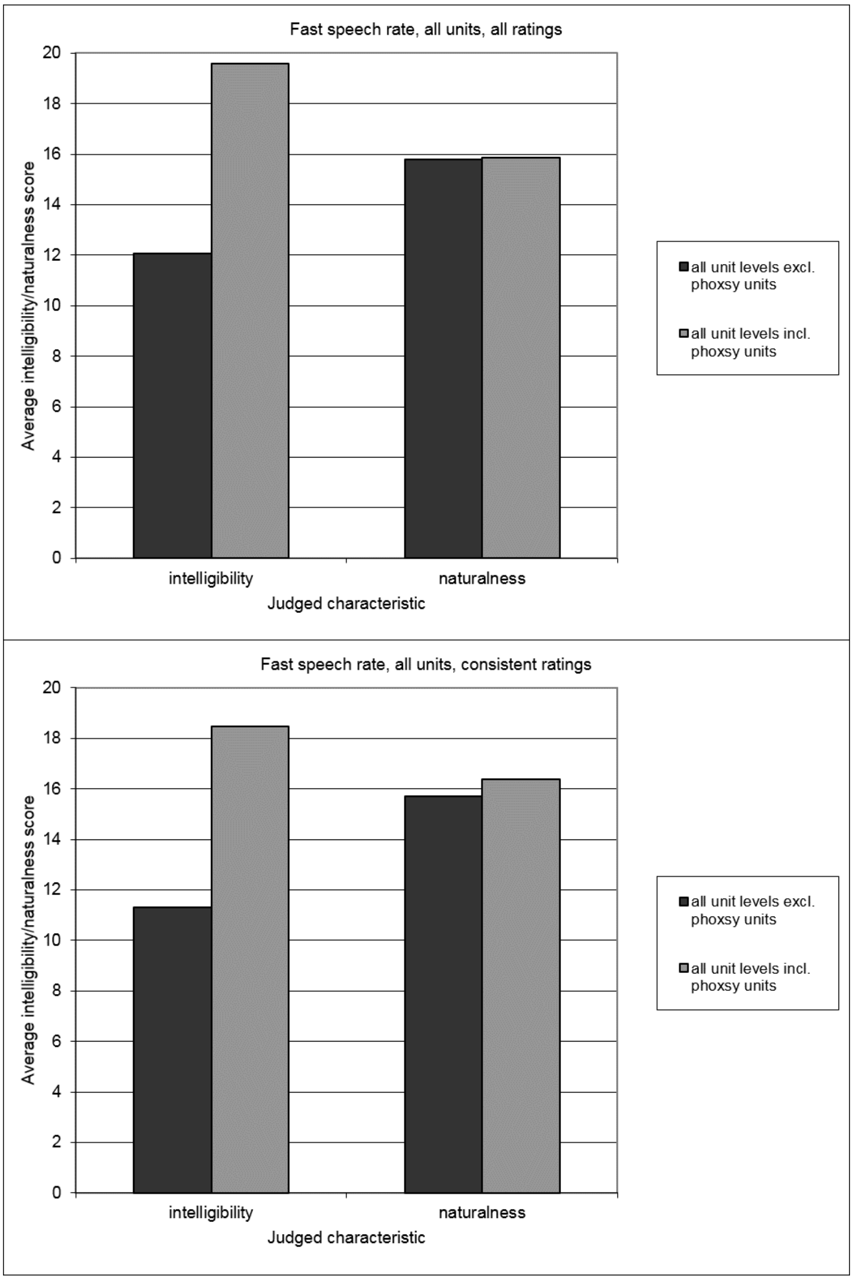


Figure 8.4: Average intelligibility and naturalness scores for fast speech stimuli generated using all units levels excluding phoxsy units (dark grey columns) or including phoxsy units (light grey columns). Top: all ratings, bottom: consistent ratings.

when proposing a new method to evaluate TTS output at higher levels of linguistic organization instead of segmental level, stated that "[i]n any case, the results of the present study suggest that comprehensibility of natural and synthesized texts does not have to be tested separately for sighted and non-sighted people, results for the one group being generalizable to the other.", Mean Opinion Scores (MOS) and Word Error Rates (WER) were collected from trained blind or visually impaired subjects as well as naive untrained subjects (mostly sighted) separately. It was anticipated that ultra-fast stimuli would be judged better by trained listeners than by untrained listeners. For the latter, it was also expected that stimuli generated from the fast speech unit selection inventory would get significantly higher scores compared to stimuli generated from the normal speech unit selection corpus as well as compared to stimuli generated by means of formant synthesis. This assumption was based on the expectation that the intelligibility of the fast speech unit selection stimuli would be comparable to the intelligibility of the other two stimuli groups, but that they would sound more natural due to less durational manipulation necessary to achieve higher speaking rates. Additionally, it was hypothesized that trained blind or visually impaired listeners would judge stimuli generated by means of formant synthesis generally better than stimuli generated from either unit selection corpus because of their habituation to this kind of speech synthesis [Arons, 1992], [Nygaard et al., 1994], [Tucker and Whittaker, 2006], [Syrdal et al., 2012]; cf. also section 4.2.3. The observations made regarding the influence of being part of one of the user groups on judgments will be discussed in more detail in connection with the results of the perceptual evaluation of synthesized fast speech elaborated in chapter 4.3.

[Jongenburger and van Bezooijen, 1992] as well as [Chalamandaris et al., 2010], [Papadopoulos et al., 2010] and [McCarthy et al., 2013] also evaluated the acceptability of different aspects of synthetic speech as a function of experience. [Jongenburger and van Bezooijen, 1992] were especially interested in an answer on the question whether experience with a certain system enhanced or inhibited the evaluation skills of the listener, and whether a carry-over effect to other kinds of (synthetic) speech existed. Acceptability was evaluated in terms of ten different criteria. However, results were rather redundant in showing similar patterns of significant effects. Already here, intelligibility and naturalness excelled as best fit to two groups of criteria (cf. [Chalamandaris et al., 2010], [McCarthy et al., 2013]). The observation that exposure to high-quality synthesized speech did not raise perceived intelligibility whereas exposure to lower-quality synthesized speech did, led [Jongenburger and van Bezooijen, 1992] to the conclusion that listeners were indeed able to learn interpret segmental characteristics of a particular speech synthesis system. However, repeated exposure did not lead to a more positive perception nor a better rating, in contrast to intelligibility judgments. This aspect of fast (synthesized) speech perception will also be discussed in chapter 4.3.

8.2.1 Perceptual Evaluation of Speech Synthesized at Ultra-Fast Speech Tempo

Methods

By means of both underlying unit selection corpora separately, twenty SUS [Benoit and Grice, 1996], cf. also [Schweitzer et al., 2004] were generated with the BOSS system [Klabbers et al., 2001]. Each SUS comprised six to seven short common words. All SUS generated are listed in appendix E. Similar to the method applied for the evaluation of corpus recordings described in chapter 7.1.1, the synthesized SUS were linearly accelerated to three (in case of the stimuli based on the fast inventory) or four (in case of the stimuli based on the normal rate inventory) different speaking rate levels by means of the TD-PSOLA implementation available in Praat [Boersma and Weenink, 2010]. The lowest speaking rate generated was four syllables per second, whereas the highest was twenty syllables per second, with equal distances of four syllables per second in between. However, the stimuli based on the fast speech corpus were not slowed down to meet the normal speaking rate of four syllables per second as this was assumed useless in the given analysis. Therefore, in the normal rate condition there were only two kinds of stimuli to evaluate: Utterances generated by means of the normal rate unit selection corpus and utterances based on formant synthesis. To generate the SUS to be evaluated also by means of formant synthesis, the popular formant synthesis system "JAWS Eloquence" [FreedomScientific, 2011] (cf. chapter 5) was used. SInce JAWS does not allow for adjusting the speaking rate in syllables per second, several utterances were generated at different speed levels whose speaking rate then was calculated from the resulting signal in syllables per second until the speaking rates desired for the current experiment were determined.

As an evaluation of all synthesized stimuli would have exceeded a reasonable amount of judgments, the experiment was created in a block design where each speaking rate level consisted of two SUS based on the normal rate corpus and two SUS based on the fast rate corpus. Additionally, an equal distribution of different stimuli across all speaking rates was implemented to minimize textual influences. This also held for the two SUS per speaking rate level generated by means of formant synthesis. The experiment was conducted in a quiet environment and stimuli were presented via earphones. It was implemented using the Praat ExperimentMFC environment [Boersma and Weenink, 2010] and consisted of six subsets of stimuli, one for unit selection based stimuli for each speaking rate condition, and one for the stimuli based on formant synthesis. The supervisor was the person to operate the experiment, thus subjects could concentrate on the listening task. Especially for the blind and visually impaired listeners searching for a play button on the computer desktop would have introduced a huge distraction from the actual task as they would have apllied a screenreader software for it.

To familiarize the listeners to the task, stimuli from the normal rate unit selection condition (four syllables per second) were played at first. Those stimuli consisted of utterances based on the normal rate unit selection inventory only. Subjects were asked to listen to the SUS presented and repeat aloud after it was played. In case the listener did not understand what was said the utterance was played again by the supervisor. Only one replay was allowed. After playing the sentence again, subjects were asked to judge overall acceptability of the presented utterance on a scale from 1 = *poor* to 5 = *excellent*. Moreover, they were advised to also include the parameters ”naturalness” and ”voice quality” into their judgment, and to not only rely on ”intelligibility”. The execution of a replay as well as the number of correctly understood words and the assigned MOS were documented by the supervisor before playing the next stimulus. This way, a baseline score for intelligibility and acceptability of unit selection stimuli based on the normal rate inventory was gained.

The following subset of stimuli presented consisted of SUS reflecting the next higher speaking rate (eight syllables per second). From this speaking rate level on, subjects were presented with two SUS based on the normal rate inventory and two SUS based on the fast rate unit selection inventory. The four utterances were played at random order. Again, listeners were asked to repeat the SUS aloud after it was presented. In case the stimulus was not understood by the listener, it was played once again. After repeating the sentence aloud, listeners judged overall acceptability of the presented utterance on a scale from 1 = *poor* to 5 = *excellent* again. Also here, they were advised to include ”naturalness” and ”voice quality” into their judgment, and to not only rely on ”intelligibility”. The execution of a replay and the number of correctly understood words as well as the assigned MOS were documented again before the next stimulus was played. This procedure was repeated for all speaking rate levels. Only afterwards, utterances generated with formant synthesis were presented. Here, each speaking rate level consisted of two SUS featured the same way as the stimuli based on unit selection synthesis were before. The procedures remained the same as in the first subsets of the experiment [Moers, 2011].

The group of blind or visually impaired listeners consisted of twenty-one people, twenty-four years old on average with 38% female subjects. All participants relied on speech synthesis applications when working with a computer and had at least two years of experience in doing so (*trained* listeners). The second group of participants consisted of seventeen sighted subjects, thirty years old on average with 35% females. None of them was experienced in using speech synthesis applications at all (*untrained* listeners).

For statistical analysis of judgments gathered, binary decision trees were applied as they allowed for taking into account several variables independent of their scaling. Variable values were recorded for each stimulus separately and merged into a database. Possible values of single variables are listed in table 8.2. Decision tree calculation was conducted by means of the *ctree* function of

Variable	Value
Listener	1 = untrained 2 = trained
Inventory	1 = normal speech rate unit selection inventory 2 = fast speech rate unit selection inventory 3 = formant synthesis
Speaking rate	1 = 4 syllables per second 2 = 8 syllables per second 3 = 12 syllables per second 4 = 16 syllables per second 5 = 20 syllables per second
Replay	1 = no 2 = yes
Number of intelligible words	1 = 1 2 = 2 3 = 3 4 = 4 5 = 5 6 = 6 7 = 7
Mean Opinion Score	1 = excellent 2 = good 3 = satisfactory 4 = pass 5 = inadequate

Table 8.2: Variables and possible values.

the statistical software R [RDevelopmentCoreTeam, 2011]. Additionally, the WER was calculated as percentage of wrongly understood words compared to all words (cf. [Dupoux and Green, 1996]). As it was expected that results would show significant differences between listener groups regarding synthesis technique and inventory as well as WER (in terms of number of intelligible words), analyses were conducted for each listener group separately at first. Only afterwards, the outcome of the experiment was evaluated for both listener groups together. Accordingly, results are depicted separately for each listener group first, before overall findings are described in detail in the following.

Results: Untrained listeners

Figure 8.5 shows the decision tree displaying judgments made by untrained listeners dependent on the applied unit selection inventory (Inv), the number of correctly understood words (NumW), the speaking rate (SR), and the execution of a replay (Rep.). The number of correctly understood words is the first split criterion here (node 1, $p<.001$). This shows that even for untrained listeners the intelligibility is the most important characteristic of synthetic speech in general; naturalness and voice quality are not the first characteristic to pay attention to. On this upper level of the decision tree, the distinction is made between the number of words correctly understood being equal or less than two on the left hand side, and the number of words correctly understood being more than two on the right hand side. For the number of words correctly understood being equal or less than two (left hand side), the next split criterion again is the number of words correctly understood (node 2, $p<.001$). If no words were intelligible, the MOS was significantly worse compared to utterances where one or two words were understood. This observation supports the view that intelligibility is the foremost judging criterion [Chalamandaris et al., 2010], [McCarthy et al., 2013].

Looking at the right side of the tree, another variable is used as next split criterion: the execution of a replay. As the execution of a replay was restricted to one, it is not surprising that it is the split criterion here (node 5, $p<.001$). For stimuli which were immediately intelligible, the MOS is significantly better than for stimuli which had to be replayed. Correlating the number of correctly understood words with the MOS shows a significant correlation ($r=-0.662$) as well. The same holds for the relation between MOS and speaking rate ($r=0.558$). As the value of the latter is smaller than the value of the first correlation, it does not provide additional statistically relevant information, and therefore is not taken into account when the decision tree is calculated. Contrary to trained listeners, there is no significant difference in MOS for any of the unit selection inventories versus formant synthesis (cf. section 8.2.1). The initial assumption that this may be related to the general lack of experience with speech synthesis had to be revised when both listener groups were compared and variables taken into consideration for decision tree calculation were restricted to a subset (cf. section 8.2.1, [Moers, 2011]).

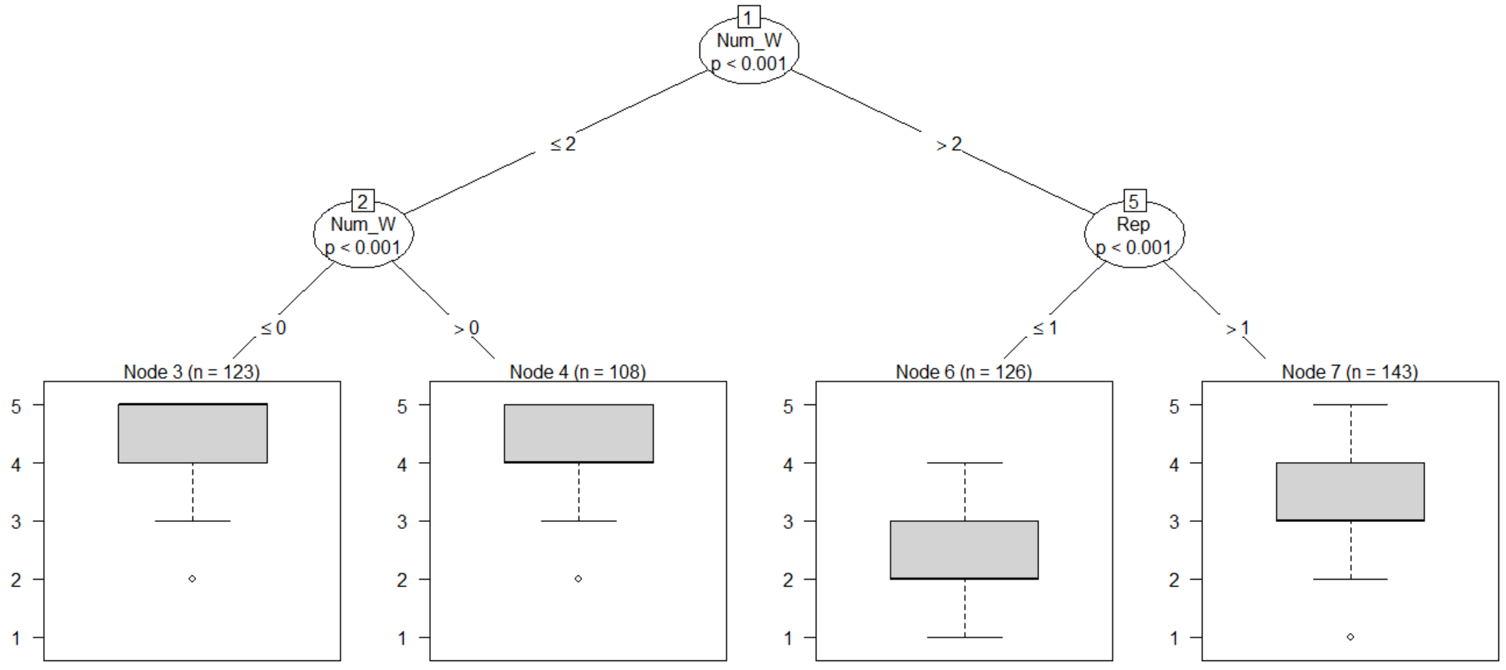


Figure 8.5: Decision tree to predict judgments by untrained listeners taking into account the number of intelligible words (NumW), the speaking rate (SR), the inventory (Inv), and the execution of a replay (Rep.).

Results: Trained listeners

Similar to figure 8.5 showing the dependencies of judgments on variables evaluated for untrained listeners, figure 8.6 shows the decision tree depicting the judgments made by trained listeners dependent on the inventory (Inv), the number of correctly understood words (NumW), the speaking rate (SR), and the execution of a replay (Rep.). Also for this listener group the number of intelligible words is the first split criterion (node 1, $p<.001$), as was expected. However, in contrast to untrained listeners, here the distinction is made between the number of words correctly understood being equal or less than four on the left hand side, and the number of words correctly understood being more than four on the right hand side of the tree. For the number of intelligible words being equal or less than four (left hand side), the next split criterion again is the number of intelligible words (node 2, $p<.001$). If no words were understood, the MOS is significantly worse than for utterances where one to four words were understood. This again supports the view that intelligibility was the foremost judging criterion for trained listeners, even more than for untrained listeners, since for the latter the split between significantly higher or lower scoring was already made for more or less than two intelligible words.

On the right hand side of the tree, another variable is used as the next split criterion again: the execution of a replay. Whether or not the stimulus was replayed results in a significant difference in the MOS (Node 5, $p<.001$). For stimuli which were immediately intelligible, the MOS is significantly better than for stimuli which had to be replayed. As opposed to the decision tree for the untrained listeners, another split follows on the left side as well as on the right side of this part of the decision tree. On the left hand side, which depicts the partial tree for stimuli which had not to be repeated, a significant difference is found for the underlying inventory (node 6, $p<.001$). The MOS for stimuli based on formant synthesis is significantly better than the MOS for stimuli based on unit selection synthesis, no matter what speaking rate the unit selection inventory was based on. This is seen as a clear indication of a training effect for listeners used to listen to speech output generated with formant synthesis, especially with JAWS, and is in line with findings of e.g. [Winters and Pisoni, 2004], [Höpfner, 2008], and [Stent et al., 2011] who ascribe this phenomenon to the "familiarity with a synthesizer". The training effect results in a higher intelligibility of formant synthesis in general, and of JAWS Eloquence in particular, and therefore in a better MOS; naturalness and voice quality do not play any role in this judgment. In contrast, if the stimulus had to be replayed the split criterion again was the number of intelligible words (node 9, $p<.05$). When the number of intelligible words was five or less, the MOS was significantly lower than for utterances where more than five words were understood correctly. Correlating the number of intelligible words with the MOS assigned showed a highly significant correlation ($r=0.729$) as well. The same holds for the relation between MOS and speaking rate ($r=0.565$). Since the value of the latter is smaller than the value of the first correlation it

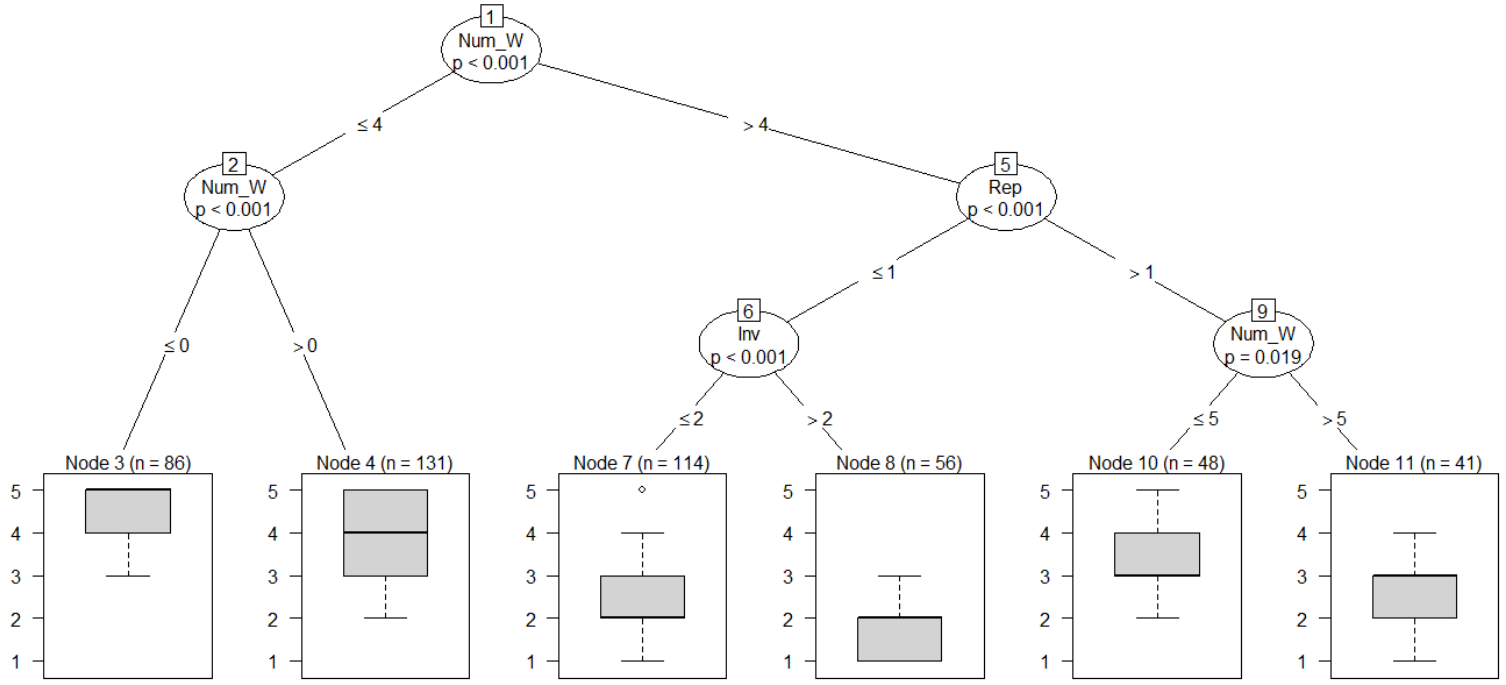


Figure 8.6: Decision tree to predict judgments by trained listeners taking into account the number of intelligible words (NumW), the speaking rate (SR), the inventory (Inv), and the execution of a replay (Rep.).

does not provide additional statistically relevant information, and therefore is not taken into account when the decision tree is calculated [Moers, 2011].

Comparing listener groups

Figure 8.7 depicts the decision tree showing judgments of synthesized speech dependent on listener group (List), inventory (Inv), number of correctly understood words (NumW), speaking rate (SR), and execution of a replay (Rep.). Again, the number of intelligible words is the first split criterion (node 1, p<.001). It was concluded that also for both listener groups taken together, the intelligibility was the most important characteristic of the generated synthetic speech in general; naturalness and voice quality did not play any important role. For this decision tree, the first distinction is made between the number of intelligible words being equal or less than four on the left hand side, and the number of intelligible words being more than four on the right hand side of the tree. For the number of words correctly understood being equal or less than four (left hand side), the next split criterion again is the number of words correctly understood (node 2, p<.001). If one or none word was intelligible, the MOS was significantly worse than for utterances where two to four words were understood. Staying on the left side of this subtree, the number of intelligible words is taken as a split criterion a third time (node 3, p<.001). Utterances which were completely unintelligible were judged significantly worse than utterances of which at least one word was understood correctly. Looking at node 6 at the right side of the left subtree, it gets obvious that again the execution of a replay induces a significant difference in the MOS (node 6, p<.001). Going further to the right, the underlying inventory causes a significant difference in judgments only for utterances of which two to four words were correctly understood (node 8, p<.05).

Turning from the top of the tree to the right hand side, it becomes clear that also for utterances for which more than four words were understood correctly, the execution of a replay caused a significant difference in judgments (node 11, p<.001). Stimuli with more than four intelligible words and no replay were judged significantly better than stimuli containing more than four intelligible words which had to be replayed. For the latter, again the number of correctly understood words was applied as a split criterion (node 13, p<.05). If more than five words were intelligible, a stimulus was judged significantly better than a stimulus of which two to five words were understood correctly. Surprisingly, neither listener group nor speaking rate seemed to have a significant influence on the MOS assigned by both listener groups.

Since no significant influence of speaking rate nor listener group was found in the first analysis, it was decided to leave out the two variables which were used as a split criterion most frequently - number of intelligible words and execution of a replay - from the next analysis step [Moers, 2011]. A new decision tree was calculated taking into account only the inventory (Inv), the speaking rate (SR), and the listener group (List). Results are depicted in figure

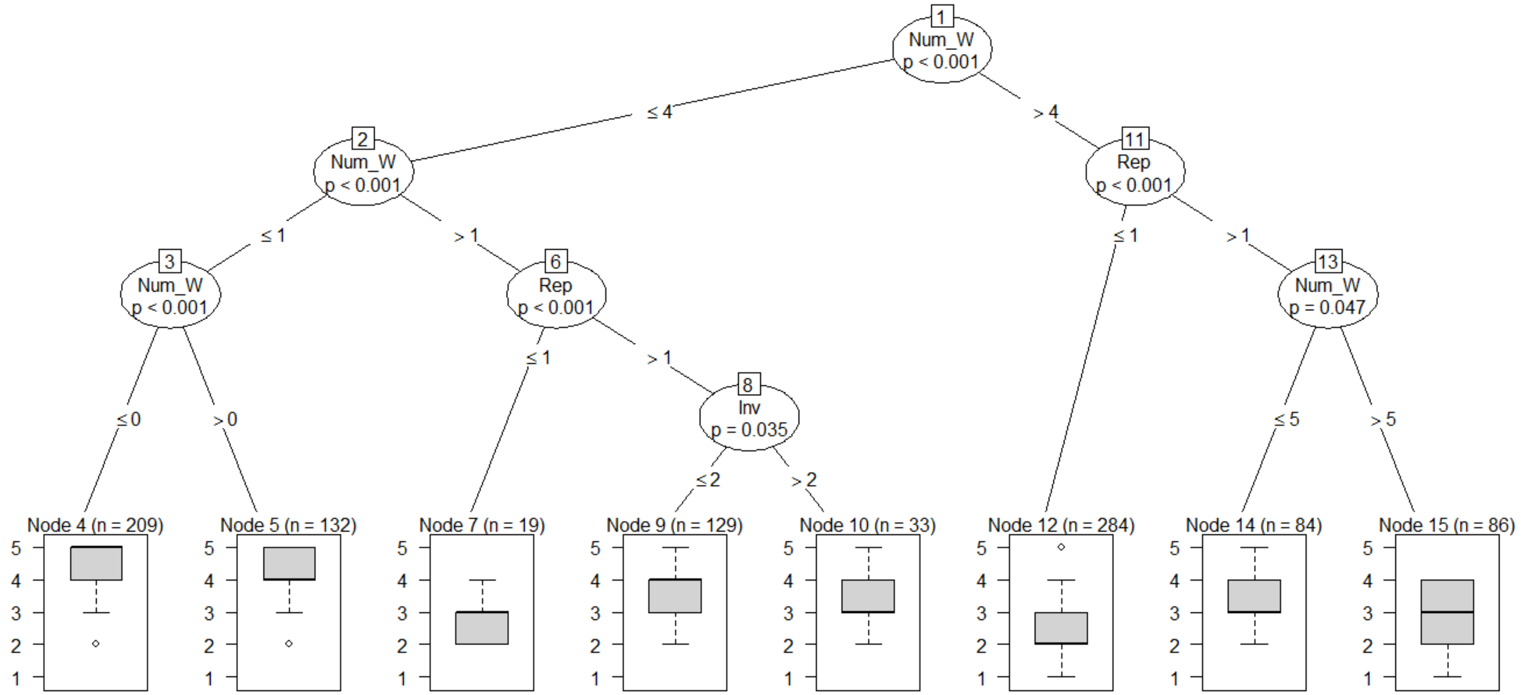


Figure 8.7: Decision tree to predict judgments by both listener groups taking into account the number of intelligible words (NumW), the speaking rate (SR), the inventory (Inv), the number of repetitions (Rep.), and the listener group itself (List).

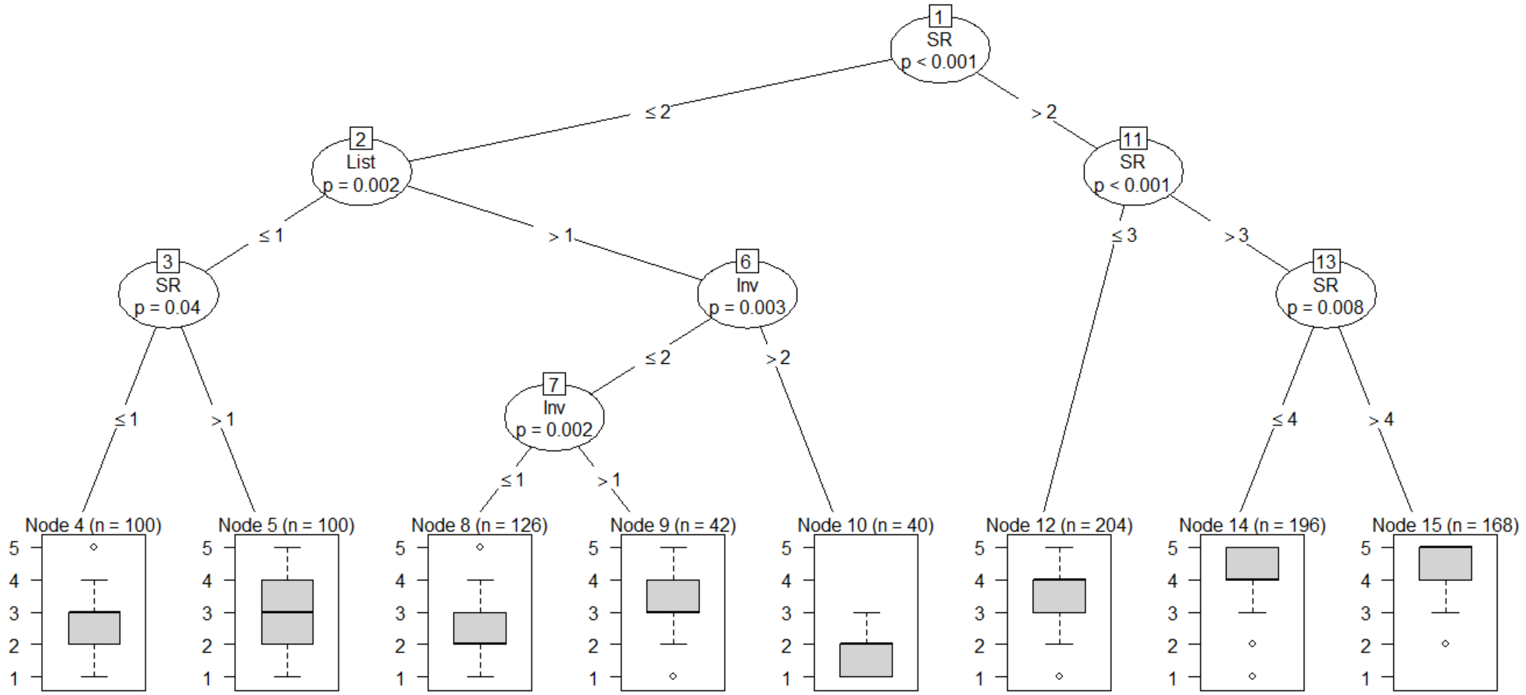


Figure 8.8: Decision tree to predict judgments by both listener groups taking into account the speaking rate (SR), the inventory (Inv), and the listener group itself (List).

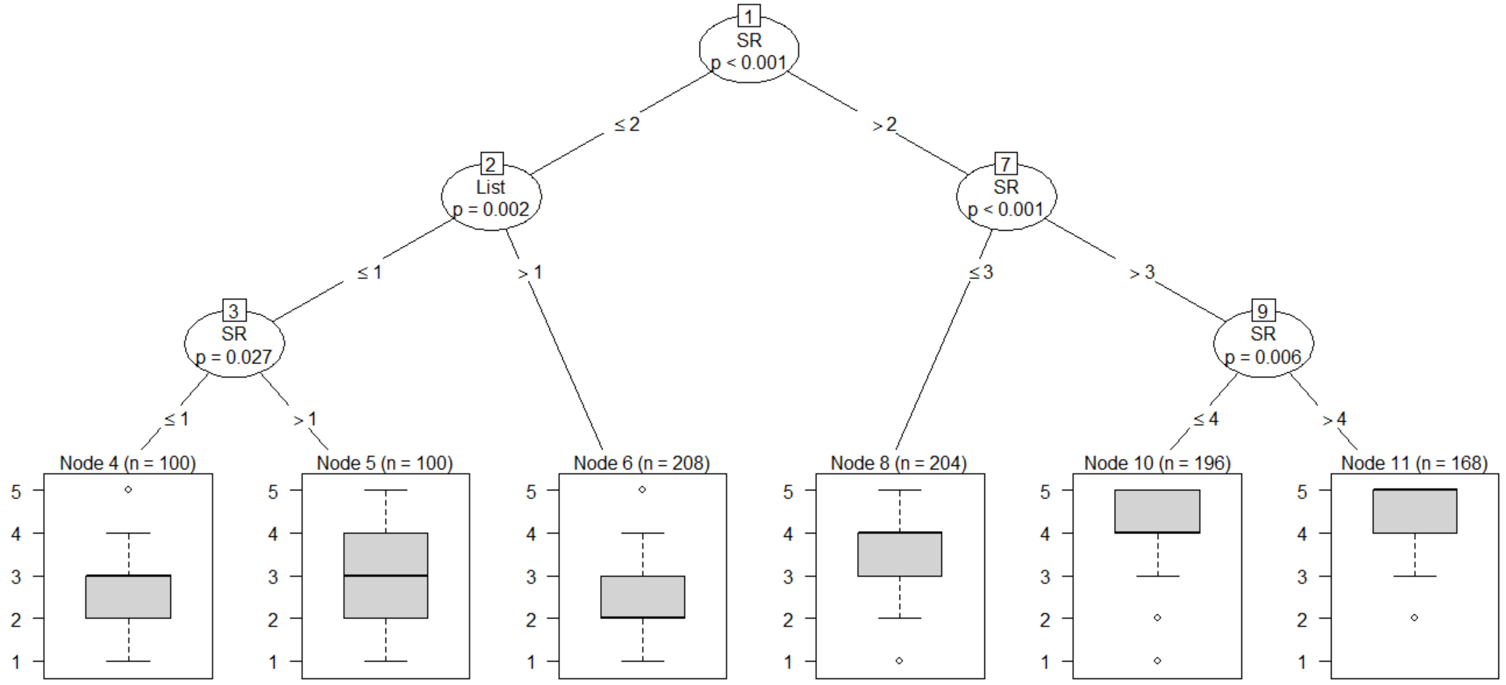


Figure 8.9: Decision tree to predict judgments by both listener groups taking into account the speaking rate (SR) and the listener group itself (List).

8.8. This approach finally showed that speaking rate as well as listener group as well as inventory played an important role in rating the stimuli presented during the experiment. Speaking rate here is the first split criterion (node 1, p<.001). Interestingly, this topmost split reveals that there is a highly significant difference in judgments between stimuli generated at a speaking rate also achievable in natural fast speech production (SR<=2, being equivalent to SR<=8 syllables per second), and stimuli generated at a higher and therefore unnatural speaking rate (SR>2), in accordance with findings by [Dietrich et al., 2013]. Following the right branch of the tree, the next split criterion again is the speaking rate (node 11, p<.001), and also the next split is based on the speaking rate (node 13, p<.01). This leads to the conclusion that the faster the speaking rate the worse the MOS, independent of listener group or inventory. A highly significant correlation (r=-0.771) between number of intelligible words and speaking rate confirms this finding. The decision tree displayed in figure 8.9 which was calculated taking into account only the variables speaking rate (SR) and listener group (List) illustrates this fact more precisely.

Looking at the left side of the decision tree depicted in figure 8.8 showing the judgments for stimuli generated at a speaking rate naturally achievable, significant differences in MOS also show up for different listener groups and underlying inventory: The first distinction made after the topmost split here is for listener group. Untrained listeners judged stimuli significantly worse than trained listeners in general (node 2, p<.01). This again can be interpreted as evidence for a training effect regarding speech synthesis occurring for frequent and therefore trained listeners. Hereafter, the following split criterion for the group of untrained listeners again was the speaking rate (node 3, p<.05); stimuli generated at a normal speaking rate (four syllables per second) were scored significantly better than stimuli generated at a faster speaking rate (eight syllables per second). Note that no distinction was made for the underlying inventory here. For trained listeners, however, the next significant difference in judgments was found for the inventory. Stimuli genrated by means of formant synthesis were judged significantly better than stimuli generated by means of unit selection synthesis (node 6, p<.05), and within the unit selection based stimuli, utterances based on the normal speech rate inventory were judged significantly better than utterances based on the fast and clear speech inventory (node 7, p<.05). This result is depicted in more detail in figure 8.10 as well. Leaving out all variables other than inventory and listener group, the resulting decision tree clearly indicates that trained listeners were biased towards formant synthesis (node 9). For both listener groups, judgments for stimuli based on the normal speech unit selection inventory were significantly better than for utterances generated by means of the fast and clear speech inventory (node 2, node 6). However, untrained listeners did not distinguish between fast speech generated by means of unit selection synthesis and fast speech based on formant synthesis; the decisive criterion here is only the speaking rate again [Moers, 2011].

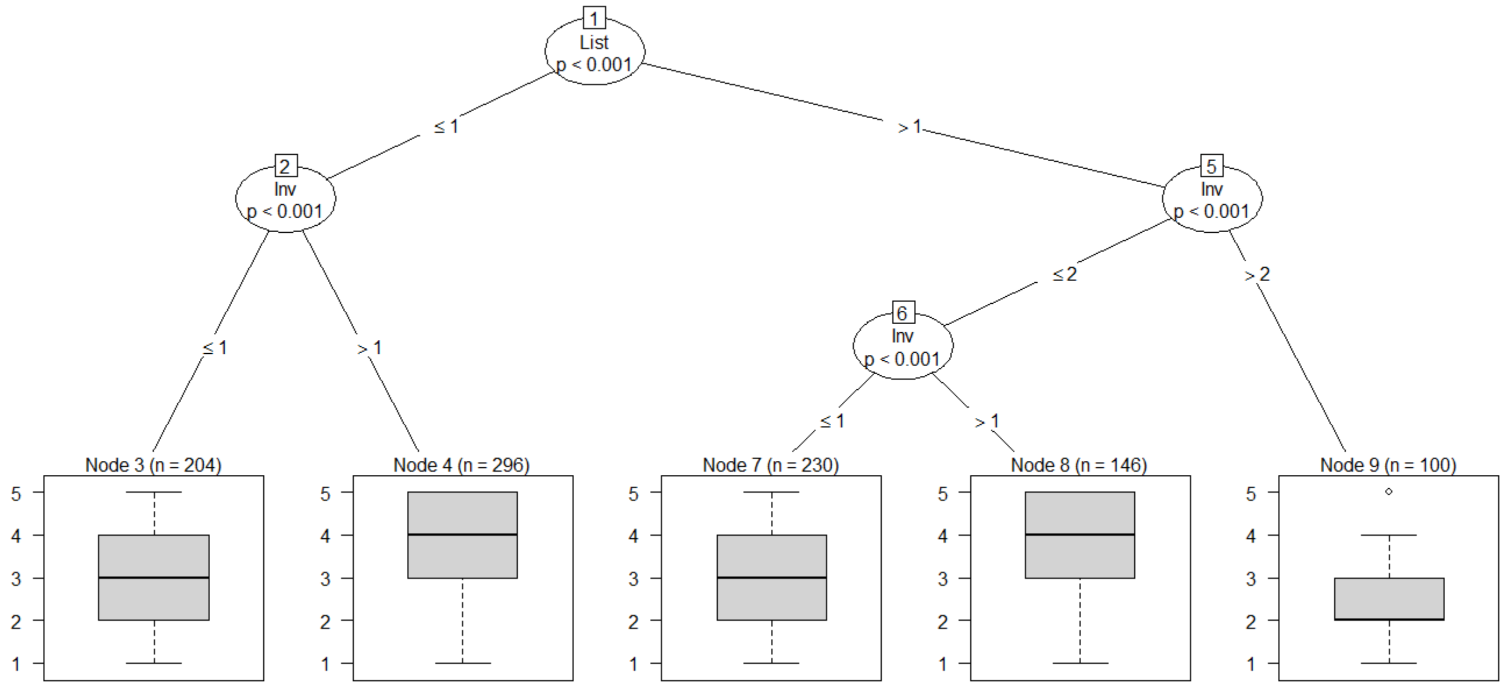


Figure 8.10: Decision tree to predict judgments by both listener groups taking into account the inventory (Inv) and the listener group itself (List).

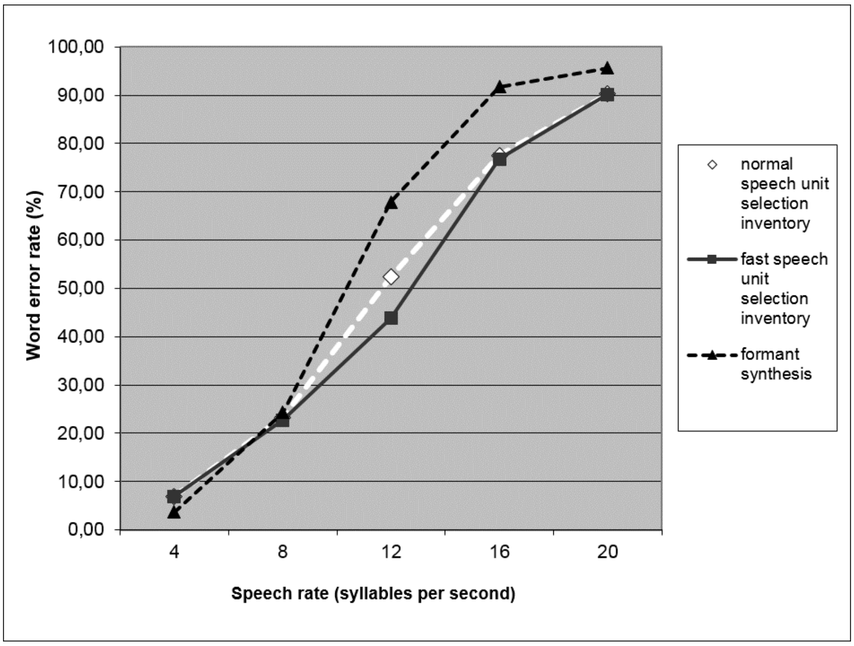


Figure 8.11: WER for untrained listeners for increasing speech rate with different synthesis approaches.

The habituation of trained listeners to formant synthesis is also reflected in the Word Error Rate (WER) for the SUS generated at different speaking rates with different synthesis approaches. The WER is a commonly used alternative format for the "number or percentage of intelligible words" [Altmann and Young, 1993] which was already introduced in chapter 4.2.1. The WER for trained and untrained listeners for both unit selection approaches as well as for formant synthesis is displayed again in figures 8.11 and 8.12 for reasons of better demonstration. In figure 8.11, the WER of untrained listeners for different synthesis approaches and increasing speaking rate is depicted. The white line refers to the unit selection inventory recorded at normal speaking rate, the grey line reflects the WER for the fast and clear speech inventory, and the dotted black line depicts the WER for formant synthesis. The WER does not differ significantly for all three synthesis approaches. This fact can be found back in the decision tree depicted in figure 8.5 where the inventory is not used as a split criterion at all. In figure 8.12, the WER of trained listeners are shown. Whereas the WER for both unit selection approaches does not differ significantly, the WER for SUS generated by means of formant synthesis is significantly lower. In figure 8.6, this circumstance is reflected in the split made in node 6.

These results show the two most important findings of the current evaluation: One is the existence of a training effect for trained listeners with regard to synthetic speech in general, and even more with a strong preference for formant synthesis in particular, reflected in significantly better judgments for

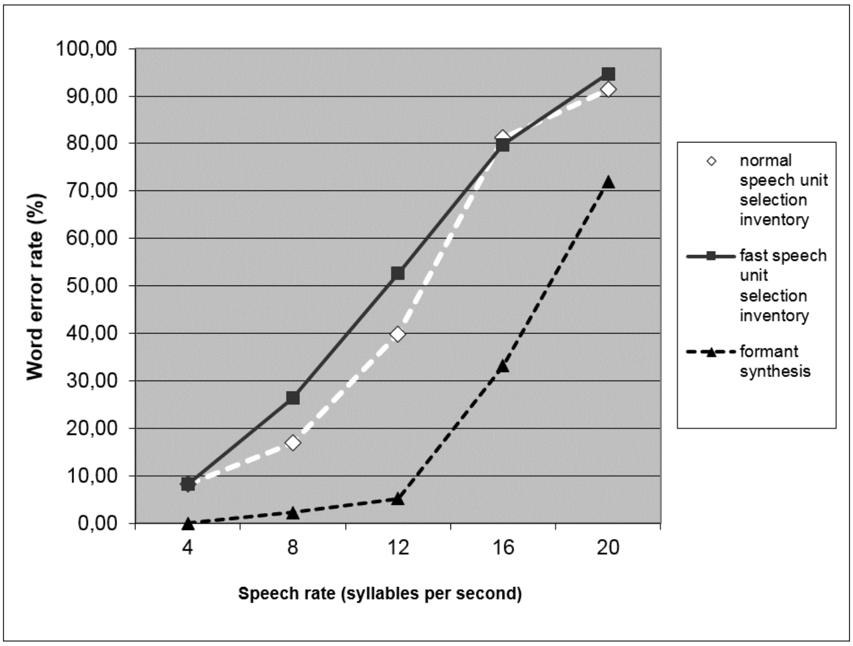


Figure 8.12: WER for trained listeners for increasing speech rate with different synthesis approaches.

utterances synthesized with this kind of speech synthesis technique compared to sentences generated by means of unit selection synthesis. The second and even more important finding is that the fast and clear speech unit selection inventory developed in the course of the work presented here did not show any advantages regarding intelligibility, naturalness, or overall acceptability in generating fast speech in unit selection synthesis, neither for untrained listeners who made no distinction between inventories for fast synthesized speech at all, nor for trained listeners who judged utterances based on the implemented fast and clear speech unit selection inventory worse than stimuli generated by means of the normal speech rate unit selection inventory [Moers, 2011].

8.3 Summary and Conclusions

To define the adequate unit to synthesize fast speech, multi-phone (”phoxsy”) units [Breuer and Abresch, 2004] were implemented as an independent multi-phone unit level in BOSS. An evaluation of speech synthesized from a normal rate speech corpus showed a significant advantage in intelligibility for stimuli generated by using only phoxsy units compared to stimuli synthesized by using only phones. For stimuli generated from all unit levels including phoxsy units the intelligibility scores were significantly higher than for stimuli generated by leaving out phoxsy units for synthesis as well. However, the level of significance was not as high as for the single unit condition. For naturalness judgments, no significant difference between the two underlying versions in both the single unit and the all unit levels condition was found. Thus, the results presented by [Breuer and Abresch, 2004] were confirmed for normal rate speech.

For fast and clear speech, the picture was strikingly different: The evaluation of utterances synthesized from a fast and clear speech inventory showed a significant advantage in both intelligibility and naturalness for stimuli generated from phoxsy units only. Stimuli generated from all unit levels including phoxsy units revealed a significant difference in intelligibility judgments as well when compared to stimuli generated from all unit levels excluding phoxsy units. Since the evaluation of stimuli synthesized from the fast speech corpus yielded more significant results than the analysis of stimuli generated from the normal speech corpus, it was concluded that phoxsy units were not only applicable to enhance the intelligibility of speech synthesized from a normal rate inventory, but to significantly improve intelligibility and naturalness of fast speech synthesized from an independent fast and clear speech inventory.

Afterwards, a perceptual evaluation of speech synthesized from the fast speech unit selection corpus compared to speech generated from the normal speech unit selection corpus as well as to speech synthesized by means of the popular formant synthesis system ”JAWS Eloquence” [FreedomScientific, 2011] was conducted. To accommodate the possible bias for trained blind and visually impaired users regarding formant synthesis, untrained people were included in the evaluation as a control group. Utterances synthesized from different underlying corpora and different speech synthesis systems at different speaking rates were evaluated with regard to intelligibility, naturalness, and overall acceptability. A MOS was collected from the two different listener groups for SUS [Benoit and Grice, 1996] generated by means of both unit selection corpora as well as with formant synthesis to investigate their intelligibility, naturalness, and overall acceptability as well. Additionally, the WER was analyzed depending on speaking rate, listener group and underlying synthesis approach.

The results of statistical analyses conducted by means of binary decision trees, taking into account all available variables, showed that for both listener groups intelligibility was the foremost judging criterion, reflected in ”number of words” being the topmost split criterion. Naturalness and voice quality did not play an important role. However, in contrast to trained listeners, MOSs collected from untrained listeners revealed significant differences for utterances of which only two or less words were understood, whereas for trained listeners this significance only appeared for utterances of which four or less words were intelligible. Additionally, for trained listeners MOS scores for stimuli based on formant synthesis were significantly better than MOS scores for stimuli based on unit selection synthesis, no matter what speaking rate the unit selection inventory was based on, whereas no significant difference in scores was found for any of the inventories or underlying synthesis systems for untrained listeners. This was seen as a clear indication of a training effect for listeners used to listen to synthesized speech, especially speech generated by means of formant synthesis [Winters and Pisoni, 2004].

Leaving out the two variables which were used as a split criterion most frequently in the first analysis finally revealed that also the variables "speaking rate", "listener group", and "inventory" played an important role for the rating of the stimuli presented. Judgments for stimuli based on the normal speech unit selection inventory were significantly better than for utterances generated by means of the fast and clear speech inventory for both listener groups. However, untrained listeners did not distinguish between fast speech generated by means of unit selection synthesis and fast speech based on formant synthesis whereas trained listeners did. Interestingly, a highly significant difference in judgments between stimuli generated at a speaking rate also achievable in natural fast speech production and stimuli generated at a higher and therefore highly unnatural speaking rate showed up. This is in line with what was reported by [Quené, 1996]: He found the highest intelligible speaking rate to be at approximately 8.64 syllables per second whereas the most comfortable listening tempo was at 6.61 syllables per second. However, this is much lower than what was reported by [Portele and Krämer, 1996] or [Fellbaum, 1996]. Moreover, also [Trouvain, 2006], [Trouvain, 2007] noted that synthetic speech generated by means of formant synthesis at speaking rates of up to 17.5 syllables per second was still comprehensible to their blind subjects which was ascribed to the intense and long-term training these subjects had undergone. For diphone synthesis, in contrast, comprehension declined for both listener groups for speaking rates faster than 7.5 syllables per second. Nevertheless, there was a difference in comprehension scores between listener groups.

In summary, the two most important findings of the current evaluation were the affirmation of the existence of a training effect for trained listeners with regard to synthetic speech in general, and a strong preference for formant synthesis in particular, which was reflected in significantly better WER and judgments for utterances synthesized by means of formant synthesis compared to sentences generated by means of unit selection synthesis. These findings are in line with the results of [Chalamandaris et al., 2010], [Papadopoulos et al., 2010], [McCarthy et al., 2013] and others. The second and even more important finding was that the developed and implemented fast and clear speech unit selection inventory did not yield any advantages regarding intelligibility, naturalness, or overall acceptability when generating fast speech in unit selection synthesis, neither for untrained listeners - who made no distinction between inventories for fast synthesized speech at all - nor for trained listeners who judged utterances based on the fast and clear speech unit selection inventory worse than stimuli generated by means of the normal speech rate unit selection inventory.

Chapter 9

Summary and conclusion

The aim of the work presented here was to determine an optimal strategy for modeling fast speech in unit selection speech synthesis to provide potential users with a more natural sounding alternative for fast speech. This specific speaking style was assumed to be preferred by the blind and visually impaired who are reliant on the use of assistive technology in their everyday life. When using screen reader applications most of the users favor formant synthesis which is able to produce synthetic speech also at fast speech rates over other speech synthesis approaches. However, the speech generated by means of formant synthesis does not sound very natural. Unit selection synthesis systems are capable of delivering more natural output, but fast speech has not been adequately implemented into such systems to date. Thus, robust guidelines for integrating fast speech as a separate speaking style into a unit selection synthesis system were to be derived.

When starting to investigate the modeling of fast speech in unit selection speech synthesis, the phonetic characteristics of natural fast speech had to be considered first. Those characteristics are quite different from the phonetic characteristics of speech produced at an average speaking rate. Phenomena like coarticulation, reduction and elision are more likely to occur when speaking rate is accelerated. Besides from single speech segments also supra-segmental characteristics like intonation and phrasing are influenced when speech is produced at a faster rate. To begin with, different aspects of natural fast speech production were outlined in chapter 2. The definition of speaking rate was examined before different approaches to measure speaking rate were explained. Afterwards, different units of measurement were discussed in section 2.1. Here, it was determined to measure the overall speaking rate of the speech material to be evaluated in the course of the work presented here in syllables per second for single utterances.

Subsequently, the manifestation of changes in speaking rate as well as their effects on different linguistic units were described in section 2.2. It was pointed out that fast speech differs significantly from speech produced at a normal speech tempo both in articulatory and acoustic characteristics. First, common phenomena like coarticulation and reduction were discussed in section 2.2.1. Afterwards, the impact of a faster speaking rate on articulation was outlined in section 2.2.2. Since articulation has to take place in a smaller time frame in fast speech, linguistic units are usually produced with more gestural overlap and acoustic interference [Davidson, 2006]. [Jannedy et al., 2010], however, observed that articulatory movement amplitudes remained still very large in fast speech for a highly trained speaker. The authors concluded that articulatory reorganization as well as speech errors were avoided by means of training of repeated patterns (cf. [Greisbach, 1992], [Liu and Zeng, 2006]). This observation was the basic principle of the procedure applied during corpus recordings outlined in chapter 7.1. To approach the fastest speaking rate possible, the speaker generally followed the strategy of repeating accelerated renditions of an utterance several times in a row. Further details of the corpus recording procedure are recapped below.

The alterations of acoustic characteristics of single speech segments as well as of transitions between them were detailed in section 2.2.2. Acoustic alterations of single segments mostly are observable in overall duration and most characteristic acoustic features, like formant frequencies for vowels or spectral moments for consonants. Implications of changes in speaking rate for larger linguistic units such as syllables, words, and phrases were then discussed in section 2.2.3. The last section 2.3 of this chapter dealt with speaking strategies commonly applied to produce different speaking styles. It was pointed out that all phenomena observable in fast speech production may lead to a loss of distinctiveness and consequently a loss of comprehension on listeners' side. However, it was shown that speakers obey certain rules in order to keep the communication chain working. Important elements of speech are less reduced than unimportant ones. With additional articulatory effort, speakers are well able to speak both fast and clear [Lindblom, 1990]. This specific speaking style, namely fast and clear speech, was finally assumed to be applicable to create a separate unit selection inventory to be used in a unit selection synthesis system.

Since the synthesis technique applied plays a crucial role depending on the goal of research, different synthesis techniques and their advantages and disadvantages were discussed subsequently in chapter 3.1. Smooth transitions required by the emergence of coarticulatory phenomena during the articulation process were shown to be crucial for the intelligibility of natural as well as synthetic speech [Martinez et al., 1997], [Winters and Pisoni, 2004]. Those are modeled best through parametric (formant) speech synthesis. Another disadvantage of unit selection speech synthesis mentioned by [Zen et al., 2007] is that the speaking style which can be synthesized is limited to the style of the speech recorded in the unit selection database, whereas statistical models used in parametric (HMM-based) synthesis only need to be trained from a database of natural speech to generate different speaking styles. Therefore, an HMM-based synthesis system offered the ability to model different speaking styles without requiring the recording and preparation of very large natural speech databases as it is required for concatenative unit selection speech synthesis. Nonetheless, concatenative unit selection speech synthesis was the method of choice for the work presented here since the generation of more natural sounding speech was the target of investigations conducted. Still, also certain advisable adaptations of the unit definition in unit selection speech synthesis were discussed and applied to avoid numerous concatenation points destroying important smooth transitions. Details of this unit definition are summarized below. For the current research, the difference between parametric (formant) synthesis, represented by the commonly used "JAWS Eloquence" application [FreedomScientific, 2011], and the concatenative unit selection synthesis system "BOSS" [Klabbers et al., 2001] was of main interest. Thus, the approach of parametric speech synthesis was outlined in section 3.1.1, followed by data-driven speech synthesis examined in chapter 3.1.2. Afterwards, the architecture of the applied concatenative unit selection speech synthesis system "BOSS" was further detailed in section 3.1.2.

In chapter 3.2, methods of modeling speaking rate in speech synthesis were examined. At first, in section 3.2.1 different approaches to duration prediction were outlined. An adequate duration prediction enhances the perceived naturalness of synthetic speech [Brinckmann and Trouvain, 2003]. In the course of time, numerous models have been developed to describe and predict the duration of speech units by taking into account various factors to different extents. Since the duration of speech segments is affected by so many different factors, the implementation of natural fast and clear speech as a unit selection corpus in speech synthesis was presumed to require an adaptation of the duration prediction module. The feasibility of doing so was examined by applying the most common and promising approach, "Classification And Regression Trees" (CART), to the normal as well as the fast and clear speech corpus recorded for the purpose of the current research. CART was seen as a promising approach to segmental duration prediction as no hand-crafting of durational rules was necessary and large datasets could be handled easily. The applied model was considering important phonetic and prosodic features influencing segmental duration. Results of CART application to the normal as well as fast and clear speech corpus are later on.

Subsequently, possibilities to manipulate speaking rate in speech synthesis were described in section 3.2.2. To date, linear duration manipulation often is the method of choice to generate fast speech, although it has some known drawbacks. "Pitch-synchronous overlap add" (PSOLA) is the most commonly used algorithm for such tasks, although the introduction of noise for an acceleration factor of two or higher is a known disadvantage. Since a natural fast and clear speech unit selection inventory already includes all segmental and suprasegmental characteristics of fast speech, the application of such an inventory is a different approach to generate fast speech at natural fast speaking rates of up to eight syllables per second by means of unit selection speech synthesis.

[Janse, 2003b], for example, compared words whose temporal structure was similar to natural fast speech to words which were generated by linear compression from normal rate speech. She observed that words generated by linear compression from normal rate speech were judged more intelligible than words mimicking natural fast speech. The less a stimulus deviated from its canonical form, the better it was understood. This finding implies that clear fast speech is preferred over slurry fast speech comprising characteristic phenomena like reduction, elision and strong coarticulation. Thus, the approach chosen to synthesize fast speech from a fast and clear speech unit selection inventory might lead to the desired improvement. Nevertheless, to generate speech at ultra-fast speaking rates as defined in chapter 2.1 (cf. [Moos and Trouvain, 2007]), TD-PSOLA was applied despite its known disadvantages, since such ultra-fast speaking rates required by certain users of speech synthesis cannot be elicited in a natural way of speech production.

In chapter 4, different aspects of fast speech perception were examined. At the beginning, approaches to natural fast speech perception as well as various models developed to describe speech perception were discussed in section 4.1.1. Afterwards, mechanisms of perceptional adjustment and compensation with regard to durational as well as spectral characteristics of (fast) speech were discussed. The explanations on natural fast speech perception were concluded with remarks on units of speech rate perception in section 4.1.2. Subsequently, the perception of artificially generated fast speech was examined in chapter 4.2. Common methods of the perceptual evaluation of artificially generated speech in general were outlined first in section 4.2.1. The specific methods chosen for the perceptual evaluations conducted in connection with the work presented here were defined. Next to judgments of intelligibility and naturalness based on "Mean Opinion Scores" (MOS) for different sets of stimuli, the "Word Error Rate" (WER) was chosen to describe the perception of natural and/or synthesized (ultra-)fast speech.

Afterwards, perceptual processes focusing on time-compressed as well as synthesized speech were outlined in sections 4.2.2 and 4.2.3. Investigations showed that the perception of synthesized (fast) speech was more difficult than the perception of natural (fast) speech [Winters and Pisoni, 2004], [Papadopoulos et al., 2010], and also more complex than the perception of (linearly) time-compressed speech [Janse, 2003b]. Moreover, synthesized speech was less intelligible than natural speech in general [Winters and Pisoni, 2004]. This phenomenon had already been observed by [Pisoni, 1981]: They found that synthetic speech required more cognitive resources revealed by longer reaction times and more numerous errors in close shadowing than natural speech. Also recall performance was worse (cf. [Luce and Pisoni, 1983], [Bailly, 2003], [Winters and Pisoni, 2004]). However, [Winters and Pisoni, 2004] concluded that "it may take longer to process synthetic speech than natural speech, but the final levels of comprehension achieved for both types of speech are ultimately equivalent." [Winters and Pisoni, 2004]. [Schwab et al., 1985] observed that synthetic speech generated by concatenation may have damaged perceptual quality by introducing discontinuities in the speech signal, but had other advantages compared to synthetic speech generated by rule because it included robust and redundant sets of perceptual cues to individual segments in the signal which was important for a better perception. They also found that synthetic speech produced by rule lacked both the rich variability and the acoustic-phonetic cue redundancies of natural speech. Also the lack of appropriate prosodic information was a disadvantage for perception. These aspects of artificial (fast) speech perception were important for the perceptual evaluation of both the natural fast corpus recordings compared to time-compressed utterances produced at normal speaking rate by the same speaker presented in chapter 7.1.1, as well as with regard to the evaluation of the synthesized (ultra-)fast speech stimuli based on different unit selection corpora as described in chapter 8.2.

To conclude the chapter about fast speech perception, observable differences between different listener groups regarding perception and basis of judgment of synthesized (fast) speech were outlined in section 4.3. The research examined here clearly showed that in general an adaptation to artificial as well as time-compressed speech takes place over time, even if it requires longer exposure to the speaking style in question than for natural fast speech [Winters and Pisoni, 2004]. Whether and how this adaptation had an influence on the quality judgments gathered in the connection of the research presented here was discussed in chapters 7.1 and 8.2 (cf. below).

Since the preferences of blind and visually impaired users regarding speaking style in speech synthesis applications had not been investigated as much in detail as it would have been desirable for designing an optimal strategy for modeling fast speech in unit selection speech synthesis, it was decided to perform a preliminary survey among the prospective users before starting the main work on implementing fast speech in a unit selection speech synthesis system. The results of the preliminary survey were outlined in chapter 5 (cf. [Moers et al., 2007]). It revealed that the blind and visually impaired who relied on speech synthesis in their everyday life often preferred a speaking rate which went far beyond what is producible in a natural way (cf. [Moos and Trouvain, 2007], [Adank and Janse, 2009]). Thus, the possibility to choose a fast speaking rate was indeed essential as reported by [Fellbaum, 1996], [Portele and Krämer, 1996], and [Asakawa et al., 2002]. However, the results of the study also indicated that more than half of the subjects held that intonation had to be sustained, and even more respondents indicated a monotonous intonation was neither desirable nor feasible. The claim of [Fellbaum, 1996] that the blind and visually impaired preferred a monotonous fast speech synthesis being prosodically relatively close to natural fast speech did not apply to this specific group of speech synthesis users in general. However, the observation that intelligibility was the most important feature when using speech synthesis devices was further supported here (cf. [Portele and Krämer, 1996], [McCarthy et al., 2013]). A third of all participants judged naturalness as not being very important, a forth would definitely do without it [Moers et al., 2007]. This again is in line with findings by [McCarthy et al., 2013] who stated that naturalness was not the most important factor for preferring a certain speech synthesis system, especially if users are advanced in using such systems. Other observations reported in section 8.2 strengthen this view as well. In contrast, 40% of the participants in the survey indicated they did not want to pass on naturalness completely. Still, it became obvious that the distortions of the speech signal occurring when using concatenative synthesis often had such a negative effect on speech intelligibility that naturalness - although higher for artificial speech generated by means of this kind of speech synthesis systems - did not play a significant role for the blind and visually impaired listeners. The overall results of the study, however, encouraged the idea to investigate the possibility of modeling fast speech in unit selection synthesis, despite disadvantages of this synthesis approach being clearly addressed.

From Lindblom's assumptions [Lindblom, 1990], outlined previously in chapter 2.3, specific requirements for the selection of a suitable speaker were defined in chapter 6.1. Afterwards, fast and clear speech produced by the selected speaker was compared to casual fast speech produced without additional articulatory effort to confirm the selected speaker's ability to produce the required speaking style. Details of the acoustic as well as perceptual evaluation of the selected speaker's speech were outlined in section 6.2. Although the analysis of specific acoustic characteristics did not reveal the desired results in the first instance (cf. section 6.2.1), results of a perceptual evaluation showed that listeners clearly preferred fast and clear speech to casual fast speech with regard to intelligibility (cf. section 6.2.2). This way, it was verified that the selected speaker was able to produce the required speaking style in an optimal way and therefore was suitable to record a fast and clear speech corpus to be used as unit selection inventory in unit selection speech synthesis.

Based on previous findings, it was then decided to create two independent, but in terms of linguistic content identical unit selection inventories: one in normal and one in fast and clear speech. It was expected that modeling fast speech in unit selection speech synthesis based on a fast and clear natural speech unit selection inventory may increase the naturalness of fast synthesized speech without harming its intelligibility. The recording procedure was described in chapter 7.1. Four hundred sentences randomly selected from the "BITS Corpus for German" [Schiel et al., 2006] were recorded for each of the two speaking style conditions. Afterwards, a perceptual evaluation of the corpus recordings was conducted (cf. chapter 7.1.1). The first step of the experiment was a preference test of accelerated normal speech utterances compared to unmanipulated fast speech utterances with regard to intelligibility and naturalness. The second part of the evaluation comprised a preference test of stimuli generated from both underlying speaking rate utterances manipulated to meet an ultra-fast speaking rate. As expected, in the first (fast) condition stimuli generated from normal speech rate recordings were judged more intelligible than natural fast ones (cf. [Janse, 2003a]). However, this advantage disappeared in the second (ultra-fast) condition. Still, with regard to naturalness scores assigned during the first as well as second part of the experiment showed that the advantage of natural fast speech stimuli was highly significant in the fast rate condition. Nevertheless, this significant difference disappeared again in the ultra-fast condition. One important factor which may have influenced the result is the extensive manipulation of the normal rate versions which may have created artifacts known to appear when using the TD-PSOLA algorithm [Moulines and Charpentier, 1990], [Quené, 2007], [Liu et al., 2008], whereas stimuli based on clearly articulated fast speech needed less manipulation and therefore were judged more natural, though not significantly. Since at the same time stimuli based on clearly articulated fast speech were assigned an intelligibility score comparable to the extensively manipulated normal speech rate utterances it was concluded that stimuli based on fast speech had a slight advantage regarding naturalness and at least no disadvantage concerning intelligibility compared to normal rate utterances and thus were applicable to be used as a separate unit selection inventory.

The preparation of the inventory is one of the most time consuming steps during the development of new corpora to be used in unit selection speech synthesis, as usually a lot of manual work is required. Therefore, to segment speech in normal speech tempo automatic labeling techniques are preferred. However, as the quality of the synthesized speech largely depends on the "Label Timing Accuracy" (LTA, [Kominek et al., 2003]), using the same segmentation algorithm based on the same canonical transcriptions for both normal and fast speech corpus recordings might result in a considerably increased amount of incorrect labels for fast speech utterances. If so, automatic phone segmentation would not be applicable to fast speech, even if it was articulated as accurately as possible. Thus, automatic segmentation of the normal as well as the fast and clear speech corpus was conducted. Actual processing steps included the adaptation of already existing transcriptions to the needs of the BOSS system, the automatic segmentation of the two different sets of corpus recordings into speech units by means of an HTK-based aligner adapted to German [Dragon, 2005], as well as the analysis of the LTA for both corpora. The results of the LTA analysis were outlined in detail in section 7.2.1. It showed only marginal differences between normal versus fast and clear speech. Given the satisfactory segmentation performance within a commonly accepted twenty millisecond tolerance interval for both speaking styles as well as the significantly shorter segment durations in fast speech, it was concluded that automatic phone segmentation is a technique applicable to speech at both normal and fast speaking rate, at least if the latter was performed with high precision and enhanced articulatory effort. When CART-based duration prediction was applied afterwards to normal as well as fast and clear speech rate utterances, results revealed that the correlation between observed and predicted duration was comparable for corpus recordings at both speech rates (cf. chapter 7.2.2). Therefore, it was concluded that also CART-based duration prediction was applicable to normal as well as fast and clear speech.

The fast and smooth acoustic transitions enhancing intelligibility in natural speech (cf. section 2.2.2) are even more important for the intelligibility of synthetic speech [Amerman and Parnell, 1981], [Janse, 2003a]. Such transitions are not treated adequately by concatenative synthesis, but can be easily modeled by formant synthesis (cf. chapter 3.1). Corresponding to this, blind listeners preferred less natural sounding formant synthesis over more natural sounding concatenative synthesis with regards to intelligibility when listening to ultra-fast speech (cf. chapter 5, [Moos and Trouvain, 2007]). Therefore, a new approach to define units for selection in heavily coarticulated contexts developed by [Breuer and Abresch, 2004] was taken up for the work presented here: Phone sequences which are prone to heavy coarticulation are treated as atomic in the sense that they are regarded as two or more phones, but one indivisible synthesis unit to minimize concatenation points. It was assumed that this new unit definition would lead to a possible solution to modeling fast synthetic speech both more naturally by using prerecorded concatenation units and more intelligibly by including typical smooth transitions in heavily coarticulated contexts. Details of the investigation of the applicability of this so called "phoxsy units" [Breuer and Abresch, 2004] to synthesize fast speech in unit selection synthesis were described in section 8.1. Phoxsy units were implemented as an independent multi-phone unit level in the BOSS system. Different groups of stimuli were generated by means of different underlying unit definitions. For normal rate speech, results regarding intelligibility preferences as presented by [Breuer and Abresch, 2004] were confirmed. For naturalness judgments, no significant difference between different stimulus versions was found. For fast speech, the picture was strikingly different: The evaluation of utterances synthesized from the fast speech inventory showed a significant advantage in both intelligibility and naturalness for stimuli generated by means of phoxsy units. Since the evaluation of stimuli synthesized from the fast speech corpus yielded more significant results than the analysis of stimuli generated from the normal speech corpus, it was concluded that phoxsy units were not only applicable to enhance the intelligibility of speech synthesized from a normal rate inventory, but improved even more the intelligibility and naturalness of fast speech synthesized from an independent fast and clear speech inventory.

After defining the adequate unit size to synthesize fast speech, a perceptual evaluation of speech generated from the fast and clear speech unit selection corpus compared to speech generated from the normal speech unit selection corpus as well as to speech synthesized with the popular formant synthesis system "JAWS Eloquence" [FreedomScientific, 2011] was conducted. To accommodate the possible bias for trained blind and visually impaired users regarding formant synthesis, untrained listeners were included in the evaluation as a control group (cf. chapter 4.3). Utterances synthesized from different underlying corpora and different speech synthesis systems at different speaking rates were evaluated with regard to intelligibility, naturalness, and overall acceptability. A Mean Opinion Score (MOS) was collected from the two different listener groups for Semantically Unpredictable Sentences (SUS, [Benoit and Grice, 1996]) generated by means of both unit selection corpora as well as "JAWS Eloquence". Additionally, the Word Error Rate (WER) was analyzed depending on speaking rate, listener group, underlying synthesis system and number of replays requested. Results of a statistical analyses conducted by means of binary decision trees taking into account all predefined variables showed that for both listener groups intelligibility was the foremost judging criterion reflected in "number of words" and "number of replays" being the topmost split criteria (cf. [Stent et al., 2011], [McCarthy et al., 2013]). Naturalness and voice quality did not play an important role. However, in contrast to trained listeners, Mean Opinion Scores collected from untrained listeners revealed significant differences for utterances of which only two or less words were understood whereas for trained listeners this significance only appeared for utterances of which four or less words were intelligible. Additionally, for trained listeners MOS for stimuli based on formant synthesis were significantly better than MOS for stimuli based on unit selection speech synthesis, no matter what speaking rate inventory the unit selection was based on, whereas no significant difference in scores was found for any of the inventories or underlying synthesis system for untrained listeners. This was seen as a first indication of the existence of a training effect for listeners used to listen to synthesized speech, especially speech generated with formant synthesis (cf. [Winters and Pisoni, 2004]). Further analyses finally showed that also "speaking rate", "listener group", and "inventory" played an important role for the rating of the utterances presented. Stimuli based on the normal speech unit selection inventory were rated significantly better than stimuli generated by means of the fast speech unit selection inventory by both listener groups. However, untrained listeners did not distinguish between fast speech generated by means of unit selection synthesis and fast speech based on formant synthesis whereas trained listeners did. This was an additional indication of the existence of a training effect for listeners used to listen to formant synthesis.

Interestingly, a highly significant difference in judgments between stimuli generated at a speaking rate also achievable in natural fast speech production (up to approximately 8 syllables per second) and stimuli generated at a higher and therefore highly unnatural speaking rate showed up. This is in line with what was reported by [Quené, 1996]: He found the highest intelligible speaking rate for untrained listeners to be at approximately 8.64 syllables per second. However, this is much less than what was reported by [Portele and Krämer, 1996] or [Fellbaum, 1996] for trained listeners. Moreover, also [Trouvain, 2006], [Trouvain, 2007] noted that synthetic speech generated by means of formant synthesis at speaking rates of up to 17.5 syllables per second was still comprehensible to their blind subjects which was ascribed to the intense and long-term training the subjects had undergone. For diphone synthesis, in contrast, comprehension declined for both listener groups for speaking rates faster than 7.5 syllables per second [Trouvain, 2006], [Trouvain, 2007]. Nevertheless, also there a difference in comprehension scores between listener groups showed up. These results were interpreted to confirm that distortions of the speech signal introduced by concatenative synthesis have such a negative effect on speech intelligibility that naturalness - although higher for speech generated by this kind of speech synthesis systems - does not play a significant role for blind and visually impaired listeners.

A phenomenon not investigated in the current research is the relation between semantic and pragmatic aspects and speaking rate. Similar to tempo changes in a musical piece it was observed by [Nooteboom and Eefting, 1994] that speakers varied their speaking rate within an utterance relative to the linguistic content. [Quené, 2007] found that the "Just Noticeable Difference" (JND) for human speech added up to 2.5% to 5% difference in speaking rate relative to fundamental frequency. Professional speakers produced a variation of speech tempo of up to 4% depending on the degree of novelty of the information in the relevant utterance. Thus, tempo changes which were above the JND threshold were most presumably relevant for communication. A speaker may express the relevance of an utterance in a greater context simply by changing the speaking rate, and listeners can interpret a change of speaking rate as an indication of the importance of what is said [Quené, 2007]. Also, [Monaghan, 2001] showed that in fast speech only the most important information remained accented. As content words tend to have a higher information load compared to function words, content words were kept more stressed and less reduced than function words in his iproduction experiments (cf. also [Schindler, 1975], [Fant et al., 1992]). This mainly applied to the vocalic part of content words (cf. [Widera and Portele, 1999], [van Bergem, 1995]). Moreover, [Janse et al., 2003] observed that in fast speech function words often were so heavily reduced that even changes in sentence level timing took place. Parts of speech with high information density were kept whereas parts of speech with low information density were shortened more or even completely left out. This observation may lead to another approach of fast speech generation on a semantic level in future: Increased temporal and acoustic reduction of function words compared to less reduction of more important content words when generating speech at a faster speaking rate.

However, in the past the implementation of semantic and pragmatic features into speech synthesis systems turned out to be more complicated than expected. [Granström, 1991], for example, evaluated the idea of keeping only important keywords to accelerate speech rate. This idea was quickly abandoned as it was not possible to automatically determine which words were important keywords, and which were not. [Koopmans-Van Beinum, 1990], on the other hand, suggested to make use of two different models in (fast) speech synthesis: One was the "reduction model" where more naturalness in synthesized speech was gained through the occasional insertion of reduced phones (cf. [Trouvain, 2002a], [Trouvain, 2002b]), the other one was the "expansion model" which focused on informatively important words by increasing acoustic contrasts when generating speech. [Tucker and Whittaker, 2006] took up this approach in their work and developed a model for "semantic compression" which was based on important or salient elements of speech. The methods the researchers applied to accelerate speech were text summarization and removal of insignificant words. A comparative perceptual evaluation revealed that users felt to have a greater understanding of utterances compressed with semantic compression. Although this approach seems promising for the current research as well, the methods of time-compressing speech applied here were based on acoustic techniques only, not taking into account more complex semantic approaches to time-compression. However, this could be a possible topic of future investigations.

While the research presented here was performed and noted down, speech synthesis technology developed further. Emanating from statistical parametric synthesis techniques like HMM-based speech synthesis, "Neural Networks" found their way into modern speech synthesis applications. Boundaries between parametric and concatenative synthesis became more fuzzy, and today architectures are sometimes even described as "model-based" versus "example-based" instead of the conventional distinction outlined previously [van den Oord et al., 2016]. In 2016, [van den Oord et al., 2016] published an article about a "Deep Neural Network" approach to speech synthesis called "WaveNet" which was used to generate raw audio waveforms from learned relations in speech. The underlying model was auto-regressive and fully probabilistic. The predictive distribution for each audio sample was dependent on all previous ones to be able "to model the long-range temporal dependencies in audio signals." [van den Oord et al., 2016]. Nonetheless, according to the authors it was possible to train the model with a huge amount of data, for example several thousands of samples per second of audio, in an efficient way. In a perceptual evaluation of the speech generated by means of WaveNet trained on linguistic features, speech quality was rated "significantly more natural sounding than the best parametric and concatenative systems [although it] sometimes [...] had unnatural prosody by stressing wrong words in a sentence." [van den Oord et al., 2016]. Also a change in voice characteristics does not pose a problem for WaveNet as it can be conditioned to many different speakers. Thus, it seems also applicable to produce different speaking styles and fast speaking rates at a higher quality than conventional speech synthesis approaches. Nevertheless, WaveNet needs a huge amount of data for training. [van den Oord et al., 2016] note that in their case the system was trained on 24.6 hours of North American English and 34.8 hours of Mandarin Chinese to gain the described perceptual advantages. To produce such an amount of natural fast and clear speech in a consistent way would be a huge challenge for every speaker because of the repetitive nature of fast and clear speech elicitation (cf. section 7.1).

Approaches similar to WaveNet were presented by [Arik et al., 2017] and [Shen et al., 2017]. [Arik et al., 2017]'s "Deep Voice" system was meant to lay "the groundwork for truly end-to-end neural speech synthesis." (ibid.). The researchers claimed their system to be "simpler and more flexible" [Arik et al., 2017] than WaveNet since they applied Neural Networks for all system components making elaborate feature engineering and broad domain knowledge unnecessary. Moreover, the system was easier to apply to new voices, domains, or other datasets than WaveNet. Still, the speech generated did not sound as natural as speech derived from human speech units although the gap was noticeably smaller than before [Arik et al., 2017]. [Shen et al., 2017] adopted Neural Network technology to generate speech directly from text. A recurrent sequence-to-sequence feature prediction network assigned mel-scale spectrograms directly to character embeddings. Afterwards, a modified WaveNet model served as a vocoder to generate waveforms from those spectrograms [Shen et al., 2017]. The researchers state that their system reaches a MOS comparable to recorded human speech and provides a "significant simplification of the WaveNet architecture." [Shen et al., 2017].

Nowadays, the techniques of "Recurrent Neural Networks" (RNN) and "Deep Learning" are also applied in unit selection speech synthesis where they are mainly used to replace costly feature extraction and cost function development [Wan et al., 2017], [Capes et al., 2017]. [Wan et al., 2017] applied an RNN model which used a "Long Short Term Memory" (LSTM, [Gers et al., 2002] after [Pollet et al., 2017]) based auto-encoder that forged linguistic and acoustic features of each unit of a unit selection speech synthesis system into a feature vector of fixed size the authors called "embedding" [Wan et al., 2017]. Target costs are described as a distance in the "embedding space" (ibid.). Thus, unit selection is facilitated and acoustic quality is enhanced while computational costs and latency are kept at a low level [Wan et al., 2017]. [Capes et al., 2017] decided to use "Deep and Recurrent Mixture Density Networks" "to predict the target and concatenation reference distributions for respective costs during unit selection." [Capes et al., 2017]. The authors claim to significantly enhance the performance of a specific commercial hybrid unit selection speech synthesis system this way. [Pollet et al., 2017] enhanced the application of RNNs as proposed by [Fernandez et al., 2015], [Merritt et al., 2016] (after [Pollet et al., 2017]) by introducing LSTM bidirectional RNNs to predict phone duration in context, speech unit encoding and frame-level log F0 information to be used as target values when searching for applicable units. This technique allowed for storing and accessing information "over long sequences of both past and future events." [Pollet et al., 2017]. The researchers pointed out that unit selection synthesis systems still are the ones mainly used in commercial applications since they still provided higher perceptual quality compared to parametric synthesis approaches, allowed for storing of predefined utterances enhancing the generated output even more, and needed smaller computational resources when integrated into a larger system resulting in higher cost-effectiveness and viability.

To sum up, the two most important findings of the work presented here were the affirmation of the existence of a training effect for daily users of assistive speech technology with regard to synthetic speech in general and a strong preference for formant synthesis in particular which was reflected in significantly better WER and MOS judgments for utterances synthesized with formant synthesis compared to sentences generated by means of unit selection synthesis. The second and even more important finding was that despite the promising intermediate results of the investigations outlined earlier the developed fast and clear speech unit selection inventory did not yield any advantages regarding intelligibility, naturalness, nor overall acceptability when it was integrated and applied in unit selection synthesis to generate fast speech, neither for untrained listeners - who made no distinction between approaches of synthesizing fast speech at all - nor for trained listeners who judged utterances based on the fast and clear speech unit selection inventory worse than stimuli generated by means of the normal speech rate unit selection inventory. Generating tfast speech by means of unit selection synthesis at a moderate fast speaking rate may become more acceptable for the blind and visually impaired, but also for other daily users of speech synthesis applications, due to the enhanced acoustic quality of the audio output. For ultra-fast speech, however, parametric synthesis may still be the method of choice since features and characteristics cannot be statistically derived from real-life data. New approaches to speech synthesis examined above are quite promising to also be applicable to generate speech in different speaking styles including different speaking rates, like the modeling of fast and clear speech, and may point out the direction of future research.

Ein jeder hat seine eigene Art, glücklich zu sein,
und niemand darf verlangen,
dass man es in der seinigen sein soll.

Heinrich von Kleist

Appendix A

Questionnaire preliminary evaluation

1. Abschnitt: Personenbezogene Angaben

1. Bitte geben Sie ihr Alter in Jahren an (Texteingabe).

2. Bitte geben Sie ihr Geschlecht an.

   (a) weiblich

   (b) männlich

3. Bitte geben Sie an, zu welcher der folgenden Gruppen Sie gehören.

   (a) nicht sehbehindert

   (b) leicht sehbehindert (zum Beispiel Brillenträger) und bei der Nutzung eines Computers auf keine weiteren Hilfsmittel angewiesen

   (c) stark sehbehindert und bei der Nutzung eines Computers auf IT-Hilfsmittel angewiesen

   (d) blind und bei der Nutzung eines Computers auf IT-Hilfsmittel angewiesen

2. Abschnitt: Programmnutzung

1. Nutzen Sie regelmäßig Programme zur Sprachausgabe oder andere IT-Hilfsmittel?

   (a) nein, eigentlich nie

   (b) nein, nur unregelmäßig

   (c) ja, regelmäßig

2. Welche IT-Hilfsmittel und Programme zur Sprachausgabe nutzen Sie hauptsächlich?

   (a) große Schrift, Lupenfunktion

(b) Braillezeile

(c) Vorleseautomat

(d) Screenreader

(e) mehrere der genannten Programme in Kombination

(f) keine

3. Nutzen Sie IT-Hilfsmittel und Programme zur Sprachausgabe beruflich oder privat?

   (a) ausschließlich beruflich

   (b) ausschließlich privat

   (c) hauptsächlich beruflich

   (d) hauptsächlich privat

   (e) sowohl beruflich als auch privat

   (f) gar nicht

4. In welchem Umfang nutzen Sie IT-Hilfsmittel und Programme zur Sprachausgabe?

   (a) mehrere Stunden täglich

   (b) ein- bis zweimal täglich

   (c) mehrmals in der Woche

   (d) mehrmals im Monat

   (e) seltener als einmal pro Monat

   (f) nie

5. Wie lange nutzen Sie schon Programme zur Sprachausgabe?

   (a) weniger als 6 Monate

   (b) zwischen 6 Monaten und 1 Jahr

   (c) 1 bis 2 Jahre

   (d) 3 bis 5 Jahre

   (e) länger

   (f) bisher noch gar nicht

3. Abschnitt: Ausgabegeschwindigkeit

1. Welches Produkt von welchem Anbieter genau nutzen Sie hauptsächlich für die Sprachausgabe? (Texteingabe, freiwillig)

2. Haben Sie die Möglichkeit, bei dem von Ihnen verwendeten Programm die Geschwindigkeit der Sprachausgabe zu steuern?

   (a) nein, habe ich nicht

   (b) ja, aber ich nutze sie nie

   (c) ja, aber ich nutze sie nur selten

   (d) ja, ich nutze diese Möglichkeit ab und zu

   (e) ja, ich nutze diese Möglichkeit häufig

   (f) nein, denn ich nutze keine Programme zur Sprachausgabe

3. Wie häufig nutzen Sie eine hohe Sprechgeschwindigkeit beziehungsweise ein deutlich schnelleres Tempo als das normale Sprechtempo, in dem Sie zum Beispiel ein Telefonat führen würden, für die Sprachausgabe?

   (a) nie, ich nutze eher eine langsame Ausgabegeschwindigkeit

   (b) nie, ich nutze eher eine normale Ausgabegeschwindigkeit

   (c) häufig, allerdings nur für bestimmte Anwendungen

   (d) immer

   (e) ich kann keine hohe Ausgabegeschwindigkeit einstellen

   (f) ich nutze überhaupt keine Programme zur Sprachausgabe

4. Für welche Art von Texten nutzen Sie eine höhere Ausgabegeschwindigkeit beziehungsweise würden Sie sie nutzen?

   (a) nur für literarische Texte

   (b) nur für Emails und Briefe

   (c) nur für Nachrichten und aktuelle Informationen

   (d) nur für Webseiten, deren Inhalt ich zum Teil schon kenne

   (e) für alle Arten

   (f) gar nicht

4. Abschnitt: Natürlichkeit

1. Wie wichtig ist es für Sie, dass eine Stimme in der Sprachausgabe so natürlich wie ein Mensch klingt?

   (a) enorm wichtig

   (b) sehr wichtig

   (c) nicht so wichtig

   (d) eher unwichtig

   (e) total unwichtig

(f) weiß ich nicht

2. Würden Sie für eine schnellere Informationsausgabe darauf verzichten, dass die Ausgabestimme insgesamt möglichst so natürlich klingt wie ein Mensch?

   (a) ja, auf jeden Fall

   (b) ja, das wäre in Ordnung

   (c) nein, nur, wenn es nicht anders geht

   (d) nein, auf gar keinen Fall

   (e) weiß ich nicht

3. Müsste Ihrer Meinung nach in einer schnellen Sprachausgabe die natürliche Sprachmelodie erhalten bleiben?

   (a) ja, auf jeden Fall

   (b) ja, das wäre besser

   (c) nein, das ist nicht nötig

   (d) nein, das stört eher beim Zuhören

   (e) weiß ich nicht

5. Abschnitt: Aspekte hoher Ausgabegeschwindigkeit

1. Ist es für Sie bei einem höheren Ausgabetempo als der "normalen" Sprechgeschwindigkeit wichtig, dass sehr deutlich gesprochen wird und so die einzelnen Laute deutlich zu hören sind?

   (a) ja, auf jeden Fall

   (b) ja, das wäre besser

   (c) nein, das ist nicht nötig

   (d) nein, das stört eher beim Zuhören

   (e) weiß ich nicht

2. Ist es für Sie bei einer höheren Sprechgeschwindigkeit wichtig, dass die Satzzeichen in der Sprachausgabe vollständig umgesetzt werden?

   (a) ja, auf jeden Fall

   (b) ja, das wäre besser

   (c) nein, das ist nicht nötig

   (d) nein, das stört eher beim Zuhören

   (e) weiß ich nicht

3. Ist es bei einer höheren Sprechgeschwindigkeit für Sie wichtig, dass alle Pausen zwischen den einzelnen Abschnitten vollständig eingehalten werden?

   (a) ja, auf jeden Fall
   (b) ja, das wäre besser
   (c) nein, das ist nicht nötig
   (d) nein, das stört eher beim Zuhören
   (e) weiß ich nicht

6. Abschnitt: Sprachmelodie und einzelne Wörter

1. Würden Sie eine monotone, eher etwas langweiliger klingende Sprachmelodie gegenüber einer natürlichen, lebhaften Sprachmelodie in schneller Sprache bevorzugen?

   (a) ja, auf jeden Fall
   (b) ja, das wäre besser
   (c) nein, das ist nicht nötig
   (d) nein, das stört eher beim Zuhören
   (e) weiß ich nicht

2. Würde es Ihnen beim Verstehen der ausgegebenen Informationen helfen, wenn inhaltlich wichtige Wörter, wie Substantive oder Verben, sich von unwichtigeren Wörtern, wie Artikel oder Präpositionen, abheben würden?

   (a) ja, auf jeden Fall
   (b) ja, das wäre besser
   (c) nein, das ist nicht nötig
   (d) nein, das stört eher beim Zuhören
   (e) weiß ich nicht

3. Was meinen Sie, auf welche Weise sich inhaltlich wichtige Wörter wie Substantive oder Verben von inhaltlich unwichtigeren Wörtern wie Artikel oder Präpositionen unterscheiden sollten?

   (a) wichtige Wörter sollten langsamer als unwichtige Wörter ausgegeben werden
   (b) wichtige Wörter sollten schneller als unwichtige Wörter ausgegeben werden
   (c) weiß ich nicht, macht keinen Unterschied

4. Was meinen Sie, auf welche Weise sich inhaltlich wichtige Wörter wie Substantive oder Verben von inhaltlich unwichtigeren Wörtern wie Artikel oder Präpositionen unterscheiden sollten?

   (a) wichtige Wörter sollten mehr betont werden als unwichtige Wörter

   (b) wichtige Wörter sollten weniger betont werden als unwichtige Wörter

   (c) weiß ich nicht, macht keinen Unterschied

5. Was meinen Sie, auf welche Weise sich inhaltlich wichtige Wörter wie Substantive oder Verben von inhaltlich unwichtigeren Wörtern wie Artikel oder Präpositionen unterscheiden sollten?

   (a) wichtige Wörter sollten lauter als unwichtige Wörter ausgegeben werden

   (b) wichtige Wörter sollten leiser als unwichtige Wörter ausgegeben werden

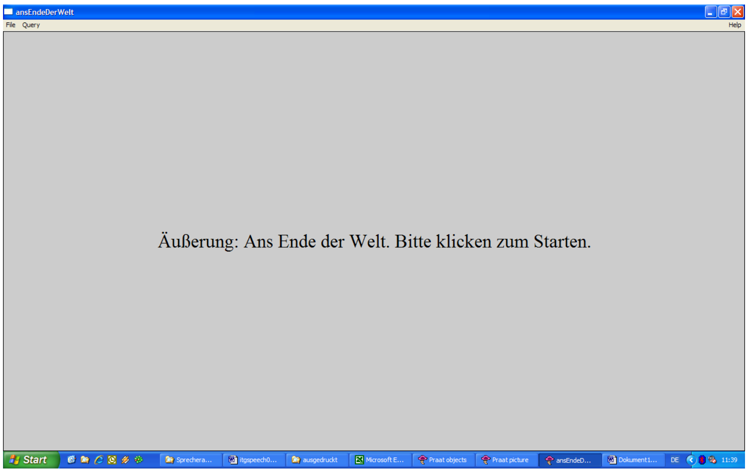
   (c) weiß ich nicht, macht keinen Unterschied

Appendix B

Setup speaker evaluation

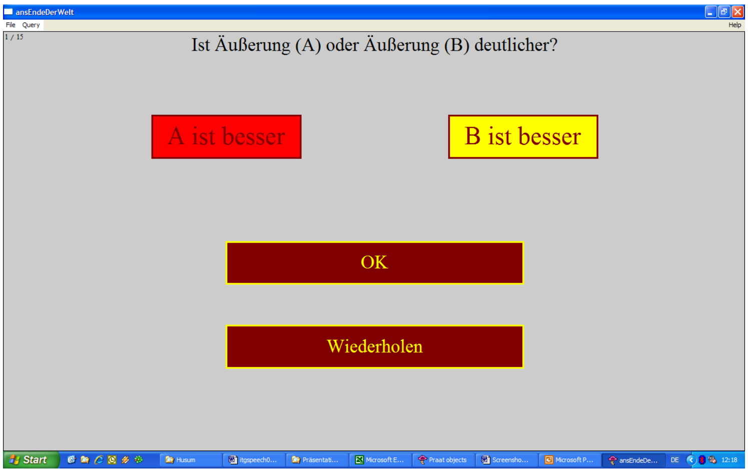
B.1 Excerpts from recorded text for speaker evaluation

- am nächsten Tag
- ans Ende der Welt
- ans Steinhuder Meer
- Berge und Wälder
- es ist eine Fahrt
- *fuhr ich nach Husum*
- hinter Gießen
- hinter Kassel die Städte
- und bei Salzgitter
- wenn bei uns Dissidenten
- *wird das Land flach und öde*



Äußerung: Ans Ende der Welt. Bitte klicken zum Starten.

Figure B.1: Text of excerpt shown before playing.



Ist Äußerung (A) oder Äußerung (B) deutlicher?

Figure B.2: Graphical user interface with option for version A or version B and replay.

Appendix C

Setup corpus recordings evaluation

- Der Schüler tippte unentwegt eine SMS auf dem Handy.
- Willst es, heb es - rennt zur Wand er, ri und rag zurück.
- Fürze werden auch als Bläherscheinungen bezeichnet.
- Mineralöl und Erdgas, deren Anteile am Gesamtverbrauch sich auf 92 Prozent belaufen.
- Damit erhöhte sie ihr Jahres-Budget für diesen Zweck auf 80.000 Euro.
- Zwei Standpunkte im Interview Ökologie kontra Ökonomie.
- Sogar beim EWR teilweise.
- Etwa 97 Prozent der Bevölkerung surft jedoch weiterhin mit dem Modem.
- In Nordfriesland gibt es jetzt einen Plattdeutsch-Beauftragten.
- Union, SPD und FDP legten gestern dafür einen gemeinsamen Gesetzentwurf vor.
- Und, weil man sich selbst der beste Freund ist.
- Fudge Tunnel, Frogs of War and That's it geben sich am Samstag die Ehre.
- Sergej Tarasenko, rechte Hand des Außenministers Schewardnadse.
- Es zog mich runter, machte mich noch nach seinem Abgang zeitweise völlig handlungsunfähig.
- Seine Doppelfunktion als CSU-Chef und Minister überfordere ihn ganz offensichtlich.
- Rechtsstaatlichkeit und Toleranz sind unteilbar.

- Zweifellos ist der Mauerpark ein Kind des Forums.
- Beim Relais erden wurde das Polyester grau.
- Wenn ein Flugzeug drauf stürzt, ist es hin.
- Hochgewachsen ist er nicht dafür drahtig, durchtrainiert und muskulös.

Appendix D

Setup unit size evaluation

D.1 Nachrichten

- Mit der Verfahrenseinstellung gegen Oberst Klein hat die Bundesanwaltschaft ihre Ermittlungen beendet.
- Der Wettbewerb des wichtigsten Filmfestivals der Welt findet 2010 ohne einen deutschen Regisseur statt.
- Ausserirdisches Leben existiere mit an Sicherheit grenzender Wahrscheinlichkeit.
- Obama schickt Raumfahrer ans Reißbrett zurück.
- Die Venus gilt als toter Planet.
- Wie liefen die Verhandlungen auf dem Weltklimagipfel in Kopenhagen wirklich ab?
- Die Mode in der DDR war viel kreativer als ihr Ruf.
- Der Verteidiger vom Deutschen Meister aus Hannover muss verletzungsbedingt passen.

D.2 Märchen

- Daheim aber stand der andere Bruder bei den Goldlilien.
- Bei Anbruch der Nacht fanden sie ein Wirtshaus und gingen hinein.
- Die Richter sprachen: bringt uns ein Wahrzeichen.
- Bei Erschaffung der Welt hatte das nächtliche Licht ausgereicht.
- Nein, wie es poltert und brummt in dem alten Elfenhügel!

- Es kam ein Soldat auf der Landstraße dahermarschiert:
- So wurde der Prinz angestellt als kaiserlicher Schweinehirt.
- Die Frau des Schusters betrachtete dieses Weib aufmerksam.
- Rosen, Tulpen, Nelken, alle Blumen welken.
- Ach, wie gut, dass niemand weiß, dass ich Rumpelstilzchen heiß!

D.3 Koran/Bibel

- Doch Jesus und seine Jünger sprachen weder Latein noch Griechisch, sondern Aramäisch.
- Wenn der Jünger ist wie sein Meister, so ist er vollkommen.

Appendix E

Setup speech synthesis evaluation

Group 1:

Subject - Verb - Adverbial: intransitive structure
Det + Noun + Verb (intr.) + Preposition + Det + Adjective + Noun

- Die Perlen stehen auf einer langen Stunde.
- Der Regen wohnt unter dem bösen Auto.
- Die Wappen danken über dem leichten Zahn.
- Der Kragen schläft in einer tollen Bank.
- Ein Ekel wartet mit der kleinen Ente.
- Das Haar weint auf den weichen Knaben.
- Ein Buch träumt unter einem dünnen Tisch.
- Die Uhr sinkt durch den roten Kittel.
- Die Nebel rennen auf das liebe Meer.
- Das Kind hilft über einer kurzen Wolke.
- Ein Kopf sitzt unter einer schweren Stiege.
- Das Gras springt mit dem kalten Teppich.
- Ein Stall liegt über der dürren Mütze.
- Die Haube reist mit der stolzen Ziege.
- Der Brei schwimmt auf der trockenen Nase.
- Ein Sarg rechnet unter einer nassen Woche.
- Der Tag schwebt über der weißen Laube.
- Eine Raupe lacht neben dem lustigen Sport.
- Eine Nonne wächst auf einem traurigen Stuhl.
- Eine Reise denkt durch den scharfen Kanal.
- Die Ohren fallen neben die bunte Maus.
- Der Onkel tanzt unter einem alten Ball.
- Ein Baum hüpft an den dicken Bus.
- Der Bauch boxt mit dem wilden Dreck.
- Ein Kamm rauscht über ein heißes Tuch.
- Die Watte erscheint durch den grauen Tee.
- Eine Gans johlt mit dem sicheren Gehirn.
- Das Sofa fliegt unter das helle Obst.

- Der Zug taucht neben eine wütende Nadel.
- Die Seite fährt mit der klammen Insel.
- Ein Rad bohrt in der dunklen Soße.
- Das Fax kräht durch einen zähen Turm.
- Ein Loch rollt mit der freien Biene.
- Der Frust spuckt durch den kühlen Faden.
- Eine Fee brummt über der vollen Scheune.
- Der Brunnen zögert neben dem krummen Käse.

Group 2:

Subject - Verb - direct Object: transitive structure

Det + Adjective + Noun + Verb (trans) + Det + Noun

- Die braune Schere sägt die Luft.
- Ein gelber Apfel bügelt die Gabel.
- Der klare Berg hebt den Wald.
- Die trübe Geige reinigt den Stern.
- Der starke Käfer streicht das Wetter.
- Eine breite Münze spielt die Wand.
- Ein grüner Winzer stört den Schrank.
- Die knappe Tasse küsst den Test.
- Der flinke Fluss liebt den Durst.
- Der laute Rahmen saniert den Gulli.
- Das tote Brot findet den Frühling.
- Die leise Pest schreibt ein Tablett.
- Die blaue Sonne reibt die Urne.
- Die prallen Ketten bedrohen den Rauch.
- Das müde Fenster klebt den Fisch.
- Der große Brief putzt das Zelt.
- Das totale öl singt eine Karte.
- Die runden Treppen schenken die Kohle.
- Das süße Licht spürt den Teich.
- Eine normale Mauer knetet den Zwerg.
- Ein eckiger Mond fährt die Butter.
- Eine schlechte Katze rührt den See.
- Die gute Birne reitet das Dach.
- Ein mürbes Blatt schlürft den Rumpf.
- Der junge Topf bewegt den Raum.
- Die flachen Würste erobern den Deckel.
- Ein hohler Kohl nimmt die Bürste.
- Der schlimme Korb schleppt einen Zoo.
- Die bärtigen Löwen bestellen den Sommer.
- Der lahme Herbst füttert den Pulli.
- Das bittere Gewehr fühlt den Saft.
- Der artige Bach füllt das Gas.
- Das freche Kamel macht den Saal.
- Eine deutsche Kerze bindet einen Tiger.
- Eine dumme Tonne fasst das Eis.
- Der schöne Typ schneidet den Trieb.

Group 3:

Verb - direct Object: imperative structure

Verb (trans.) + Det + Noun + Conjunction + Det + Noun

- Hole die Macht und die Bahn!
- Verbiete den Hals und den Hut!
- Zwinge den Mann und das Ding!
- Rette den Nabel und den Säbel!
- Beherrsche den Sinn und den Riesen!
- Miete die Wut und das Spiel!
- Knacke die Eier und die Mittel!
- Starte den Staub und die Vasen!
- Spüle den Himmel und den Mut!
- Vernichte das Maul und die Feder!
- Verpasse den Frost und das Bett!
- Klopfe den Schnee und das Glas!
- Grabe den Wagen und das Tier!
- Behaupte die Schnecke und die Frau!
- Stecke das Pferd und das Papier!
- Setze die Mitte und die Möbel!
- Biete den Keller und den Salat!
- Kassiere das Holz und die Milch!
- Versuche den Koffer und die Tulpe!
- Ernte den Traum und die Kreide!
- Zupfe die Flasche und den Winter!
- Wasche das Mehl und das Laub!
- Hacke den Reigen und die Arbeit!
- Dränge das Garn und den Fuß!
- Fälle den Essig und die Mutter!
- Danke dem Auge und dem Spaß!
- Drehe das Leben und den Platz!
- Trainiere das Geld und den Kaffee!
- Korrigiere die Wahl und den Grund!
- Leihe das Zeug und die Schule!
- Schärfe die Tasche und den Ort!
- Teste den Kerl und den Stamm!
- Schmecke den Tod und den Schatz!
- Löse den Herrn und den Job!
- Bastle die Hand und das Lob!
- Ordne die Stadt und den Honig!

Group 4:
Q.Word - Verb - Subject - direct Object: interrogative structure
Quest. Adv + Aux + Det + Noun + Verb (trans.) + Det + Adjective + Noun

- Warum stützen die Kuchen ein gemeines Teil?
- Wie raucht der Plan einen feinen Pflug?
- Wann backen die Schafe einen armen Schuss?
- Wieso sammelt die Ruhe eine brave Ecke?
- Wie schiebt der Hund ein ovales Bier?
- Wann kocht der Zucker einen blinden Affen?
- Wo lockt ein Wurm den sanften Wunsch?
- Warum schickt die Brille eine satte Hose?
- Wie legt ein Atem eine faule Musik?
- Wo tanzen die Fliegen eine fromme Schiene?
- Wieso stricken die Blicke das neue Fahrrad?
- Wann tauscht der Bär eine frische Reihe?
- Wo drucken die Häuser das hohe Ende?
- Wie grüßt die Tafel die fettigen Hühner?
- Wieso ahnt der Schluss die sauren Socken?
- Warum ärgert das Hemd einen billigen Schirm?
- Wann fängt die Tür das reiche Geschenk?
- Weshalb hält der Frosch einen ehrlichen Zaun?
- Wie regiert ein Huhn das salzige Motto?
- Wo zeichnet die Puppe das scharfe Land?
- Warum fragt die Jacke den ewigen Strauß?
- Weshalb deuten die Blumen einen ebenen Vater?
- Weshalb rollt das Regal den netten Hasen?
- Wann trinkt der Pelz ein grelles Team?
- Wie sieht das Blut die letzte Lampe?
- Wo will der Bügel den antiken Tusch?
- Wann pickt der Reis die weite Zeit?
- Warum föhnt die Nacht das seltene Bild?
- Weshalb schrubbt der Elch das schwache Los?
- Wie nennt der Schein einen rechten Mund?
- Wie teilt das Schiff die linke Erde?
- Warum schafft der Mist einen zentralen Sohn?
- Wo fährt der Film einen matten König?
- Wann nimmt der Schutz den klugen Ring?
- Warum bringen die Fragen die frühen Brüder ?
- Weshalb mag der Dank das wenige Eisen?

Group 5:
Subject - Verb - complex direct Object: relative structure
Det + Noun + Verb (trans.) + Det + Noun + Relat. Pronoun + Verb (intr.)

- Die Leute heizen den Kumpel, der gehorcht.
- Die Fahnen haben einen Rock, der spricht.
- Der Mensch baut Blusen, die schaden.
- Ein Wunder kauft eine Laus, die herrscht.
- Der Eimer meldet ein Lied, das platzt.
- Die Sachen essen eine Wonne, die vertraut.
- Eine Suppe liest eine Krise, die schweigt.
- Das Messer sucht die Rosen, die wanken.
- Der Vogel drückt einen Tenor, der lügt.
- Ein Stock beschreibt das Wasser, das winkt.
- Der Arzt weckt die Tanne, die gelingt.
- Eine Kuh malt den Witz, der siegt.
- Die Feier kaut einen Stein, der geht.
- Eine Stirn schluckt den Stift, der brennt.
- Das Herz trägt die Schuhe, die genügen.
- Das Feuer öffnet den Zwang, der jubelt.
- Ein Saum faltet ein Kino, das wackelt.
- Die Bremse tötet den Ofen, der friert.
- Die Schale jagt den Sand, der schreit.
- Eine Platte kämmt einen Besen, der humpelt.
- Ein Flug wiegt den Termin, der pfeift.
- Der Löffel flötet das Fleisch, das tippt.
- Die Schlüssel rufen die Dose, die fault.
- Der Stahl übt den Bruder, der gräbt.
- Der Freund bildet den Floh, der steigt.
- Ein Dackel glaubt den Zweig, der fehlt.
- Das Wort wählt den Krieg, der wirkt.
- Eine Welt stellt das Stück, das passt.
- Die Kraft hört einen Hunger, der regnet.
- Ein Zimmer fordert das Volk, das bleibt.
- Der Chef zeigt den Kreis, der redet.
- Das Glück stärkt einen Damm, der folgt.
- Ein Stand zählt das Büro, das droht.
- Der Weg lernt den Rest, der gefällt.
- Das Ziel schließt den Sex, der dient.
- Ein Seil schlägt das Haus, das endet.

Bibliography

[Abercrombie, 1967] Abercrombie, D. (1967). *Elements of General Phonetics.* Edinburgh University Press, Edinburgh, Scotland, UK.

[Acoustical Society of America, 2013] Acoustical Society of America (2013). Text to speech synthesis systems. http://acousticalsociety.org/search/site/S3-WG91. [Online; accessed 21-July-2017].

[Adams et al., 1993] Adams, S., Weismer, G., and Kent, R. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research*, 36:41–54.

[Adank and Janse, 2009] Adank, P. and Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of the Acoustical Society of America (JASA)*, 126:2649–2659.

[Adell et al., 2005] Adell, J., Bonafonte, A., Gomez, J., and Castro, M. (2005). Comparative study of automatic phone segmentation methods for TTS. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 999–1000, Philadelphia, PA, USA.

[Ahissar and Hochstein, 2004] Ahissar, M. and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8.10:457–464.

[Ahmed et al., 2012] Ahmed, F., Borodin, Y., Soviak, A., Islam, M., Ramakrishnan, I., and Hedgpeth, T. (2012). Accessible skimming: Faster screen reading of web pages. In *Proceedings ACM Symposium on User Interface Software and Technology (UIST)*, pages 367–378, Cambridge, MA, USA.

[Allen et al., 1987] Allen, J., Hunnicut, S., and Klatt, D. (1987). *From Text to Speech: The MITalk System.* Cambridge University Press, Cambridge, UK.

[Altmann and Young, 1993] Altmann, G. and Young, D. (1993). Factors affecting adaption to time-compressed speech. In *Proceedings Eurospeech*, pages 333–336.

[Alvarez and Huckvale, 2002] Alvarez, Y. and Huckvale, M. (2002). The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Denver, CO.

[Amano-Kusumoto and Hosom, 2010] Amano-Kusumoto, A. and Hosom, J.-P. (2010). Effect of speaking style and speaking rate on formant contours. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4202–4205, Dallas, TX, USA.

[Amerman and Parnell, 1981] Amerman, J. and Parnell, M. (1981). Influence of context and rate of speech on stop-consonant recognition. *Journal of Phonetics*, 9:323–332.

[Anward and Lindblom, 1997] Anward, J. and Lindblom, B. (1997). On the rapid perceptual processing of speech: From signal information to phonetic and grammatical knowledge. In *Presented at the International Symposium on Language Processing and Interpreting*, pages 999–1000, Stockholm, Sweden.

[Arik et al., 2017] Arik, S., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Raiman, J., Sengupta, S., and Shoeybi, M. (2017). Deep Voice: Real-time neural text-to-speech. *arXiv preprint arXiv*, 1702.07825.

[Arons, 1992] Arons, B. (1992). Techniques, perception, and applications of time-compressed speech. In *Proceedings American Voice I/O Society*, pages 169–177.

[Asakawa et al., 2002] Asakawa, C., Takagi, H., Ino, S., and Ifukube, T. (2002). The highest and the most suitable listening rate for the blind in the screen reading process. In *Proceedings of Human Interface Symposium*, pages 999–1000, Japan.

[Asakawa et al., 2003] Asakawa, C., Takagi, H., Ino, S., and Ifukube, T. (2003). Maximum listening speeds for the blind. In *Proceedings International Conference on Auditory Display (ICAD)*, pages 276–279, Boston, MA, USA.

[Aylett, 2000] Aylett, M. (2000). *Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Care of Articulation in Spontaneous Speech.* PhD thesis, University of Edinburgh.

[Bachmann and Breuer, 2007] Bachmann, A. and Breuer, S. (2007). Development of a BOSS unit selection module for tone languages. In *Proceedings 6th ISCA Speech Synthesis Workshop (SSW-6)*, Bonn, Germany.

[Bailly, 2003] Bailly, G. (2003). Close shadowing natural versus synthetic speech. *International Journal of Speech Technology*, 6.1:11–19.

[Barry, 1998] Barry, W. (1998). Time as a factor in the acoustic variation of Schwa. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 999–1000, Sydney, Australia.

[Batusek, 2002] Batusek, R. (2002). A duration model for czech text-to-speech synthesis. In *Proceedings Speech Prosody*, Aix-en-Provence, France.

[Beasley et al., 1976] Beasley, D., Maki, J., and Orchik, D. (1976). Childrens' perception of time-compressed speech on two measures of speech discrimination. *Journal of Speech and Hearing Disorders*, 41(2):216–225.

[Beaugendre, 1995] Beaugendre, F. (1995). Generating French intonation at different speaking rates. In *Proceedings Eurospeech*, pages 999–1000, Madrid, Spain.

[Bennett and Black, 2006] Bennett, C. and Black, A. (2006). The Blizzard Challenge 2006. In *Proceedings Blizzard Challenge*, pages 999–1000, Pittsburgh, PA, USA.

[Benoit et al., 1989] Benoit, C., Erp, A. V., Grice, M., Hazan, V., and Jekosch, U. (1989). Multilingual synthesiser assessment using semantically unpredictable sentences. In *Proceedings First European Conference on Speech Communication and Technology*.

[Benoit and Grice, 1996] Benoit, C. and Grice, M. (1996). The SUS test. a method for the assessment of text-to-speech intelligibility using semantically unpredictable sentences. *Speech Communication*, 18:381–392.

[Benus and Mady, 2010] Benus, S. and Mady, K. (2010). Effects of lexical stress and speech rate on the quantity and quality of Slovak vowels. In *Proceedings Speech Prosody*, volume 2, Chicago, IL.

[Beskow and Sjölander, 2000] Beskow, J. and Sjölander, K. (2000). Wavesurfer - An open source speech tool. http://www.speech.kth.se/wavesurfer/. [Online; accessed 01-March-2010].

[Birkholz, 2005] Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese*. PhD thesis, Universität Rostock.

[Birkholz, 2016] Birkholz, P. (2016). About articulatory speech synthesis. http://www.vocaltractlab.de/index.php?page=background-articulatory-synthesis. [Online; accessed 01-December-2016].

[Black and Taylor, 1994] Black, A. and Taylor, P. (1994). CHATR. a generic speech synthesis system. In *Proceedings COLING*, Kyoto, Japan.

[Black and Taylor, 1997] Black, A. and Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. In *Proceedings Eurospeech*, volume 2, pages 601–604, Rhodes, Greece.

[Black and Tokuda, 2005] Black, A. and Tokuda, K. (2005). The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. http://festvox.org/blizzard/. [Online; accessed 01-March-2008].

[Black et al., 2007] Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proceedings ICASSP*, volume 2, pages 1229–1232, Honolulu, Hawaii.

[Blauert, 1983] Blauert, J. (1983). *Spatial Hearing. The Psychophysics of Human Sound Locaization.* MIT press, Cambrige, Massachusetts.

[Boersma and Weenink, 2010] Boersma, P. and Weenink, D. (2010). Praat: Doing phonetics by computer. http://www.fon.hum.uva.nl/praat/. [Online; accessed 01-March-2010].

[Bond and Feldstein, 1982] Bond, R. and Feldstein, S. (1982). Acoustical correlates of the perception of speech rate: An experimental investigation. *Journal of Psycholinguistic Research*, 11:539–557.

[Borodin et al., 2010] Borodin, Y., Bigham, J., Dausch, G., and Ramakrishnan, I. (2010). More than meets the eye: A survey of screen-reader browsing strategies. In *Proceedings Seventh International Cross-Disciplinary Conference on Web Accessibility (W4A2010).*

[Bozkurt et al., 2003] Bozkurt, B., Ozturk, O., and Dutoit, T. (2003). Text design for TTS speech corpus building using a modified greedy selection. In *Proceedings Eurospeech*, pages 999–1000, Geneva, Switzerland.

[Bradlow, 2002] Bradlow, A. (2002). Confluent talker- and listener-oriented forces in clear speech production. In Gussenhoven, C. and Warner, N., editors, *LabPhon*, pages 241–273. Mouton de Gruyter, Berlin & New York.

[Bradlow et al., 1995] Bradlow, A., Torretta, G., and Pisoni, D. (1995). Intellgibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Progress Report No. 20 (1995)*, 20:89–115.

[Breen, 1992] Breen, A. (1992). A comparison of statistical and rule based methods of determining segmental durations. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 999–1000, Banff, Alberta, Canada.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees.* Wadsworth, Belmont.

[Breuer, 2009] Breuer, S. (2009). *Multifunktionale und multilinguale Unit-Selection-Sprachsynthese.* PhD thesis, Universität Bonn.

[Breuer and Abresch, 2004] Breuer, S. and Abresch, J. (2004). Phoxsy: Multiphone segments for unit selection speech synthesis. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 999–1000, Jeju Island, South-Korea.

[Breuer et al., 2001] Breuer, S., Abresch, J., Wagner, P., and Stöber, K. (2001). BLF - ein labelformat für die maschinelle Sprachsynthese mit BOSS II. In Hess, W. and Stöber, K., editors, *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Bonn, Germany.

[Breuer et al., 2006a] Breuer, S., Bergmann, S., Dragon, R., and Möller, S. (2006a). Set-up of a unit-selection synthesis with a prominent voice. In *Proceedings International Conference on Language Resources and Evaluation (LREC)*, pages 999–1000, Genova, Italy.

[Breuer et al., 2006b] Breuer, S., Francuzik, K., and Demenko, G. (2006b). Analysis of polish segmental duration with CART. In *Proceedings Speech Prosody*, pages 999–1000, Dresden, Germany.

[Breuer and Hess, 2010] Breuer, S. and Hess, W. (2010). The boss open synthesis system 3. *International Journal of Speech Technology*, 13:75–84.

[Breuer et al., 2005] Breuer, S., Wagner, P., Abresch, J., Bröggelwirth, J., Rohde, H., and Stöber, K. (2005). Bonn Open Synthesis System (BOSS) 3: Documentation and user manual. https://sourceforge.net/projects/boss-synth/files/BOSS%20Documentation/3.2.1/. [Online; accessed 05-November-2015].

[Brinckmann and Trouvain, 2003] Brinckmann, C. and Trouvain, J. (2003). The role of duration models and symbolic representation for timing in synthetic speech. *Journal of Speech Technology*, 6:21–31.

[Brøndsted and Printz Madsen, 1997] Brøndsted, T. and Printz Madsen, J. (1997). Analysis of speaking rate variations in stress-timed languages. In *Proceedings Eurospeech*, pages 999–1000, Rhodes, Greece.

[Bußmann, 1990] Bußmann, H. (1990). *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, Germany.

[Byrd, 1994] Byrd, D. (1994). Relations of sex and dialect to reduction. *Speech Communication*, 15:39–54.

[Byrd and Tan, 1996] Byrd, D. and Tan, C. (1996). Saying consonant clusters quickly. *Journal of Phonetics*, 24:263–282.

[Campbell, 1987] Campbell, W. (1987). A search for higher-level duration rules in a real-speech corpus. In *Proceedings Eurospeech*, pages 285–288, Edinburgh, Scotland, UK.

[Campbell, 1988a] Campbell, W. (1988a). Extracting speech-rate values from a real-speech database. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 999–1000, New York, NY, USA.

[Campbell, 1988b] Campbell, W. (1988b). Speech-rate variation and the prediction of duration. In *Proceedings COLING*, Stroudsburg, PA, USA.

[Campbell, 1990] Campbell, W. (1990). Analog I/O nets for syllable timing. *Speech Communication*, 9:57–61.

[Campbell, 1996] Campbell, W. (1996). CHATR: A high-definition speech re-sequencing system. In *Proceedings 3rd ASAASJ Joint meeting*, pages 1223–1228, Hawaii, USA.

[Campbell and Isard, 1991] Campbell, W. and Isard, S. (1991). Segment durations in a syllable frame. *Journal of Phonetics*, 19:37–47.

[Capes et al., 2017] Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., Prahallad, K., Raition, T., Rasipuram, R., Townsend, G., Williamson, B., Winarsky, D., Wu, Z., and Zhang, H. (2017). Siri on-device deep learning-guided unit selection text-to-speech system. In *Proceedings Interspeech*, pages 4011–4015, Stockholm, Sweden.

[Carlson, 1991] Carlson, R. (1991). Duration models in use. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 243–246, Aix-en-Provence, France.

[Carlson and Granström, 1975] Carlson, R. and Granström, B. (1975). Perception of segmental duration. In *Structure and process in speech perception.* Springer, Berlin, Heidelberg, Germany.

[Carlson et al., 1979] Carlson, R., Granström, B., and Klatt, D. (1979). Some notes on the perception of temporal patterns in speech. In Lindblom, B. and Öhman, S., editors, *Frontiers of Speech Communication research.* Academic Press, London, UK.

[Carlson et al., 1976] Carlson, R., Granström, B., and Larsson, K. (1976). Evaluation of a text-to-speech system as a reading machine for the blind. In *STL QPSR*, volume 2-3, pages 9–13.

[Caspers and van Heuven, 1991] Caspers, J. and van Heuven, V. (1991). Phonetic and linguistic aspects of pitch movements in fast speech Dutch. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 174–177, Aix en Provence, France.

[Caspers and van Heuven, 1992] Caspers, J. and van Heuven, V. (1992). Phonetic properties of Dutch accent lending pitch movements under time pressure. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 731–734, Banff, Alberta, Canada.

[Chalamandaris et al., 2010] Chalamandaris, A., Karabetsos, S., Tsiakoulis, P., and Raptis, S. (2010). A unit selection text-to-speech synthesis system optimized for use with screen readers. In *Proceedings IEEE Transactions on Consumer Electronics*, pages 1890–1897, Reading, UK.

[Charpentier and Stella, 1986] Charpentier, F. and Stella, M. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2015–2018, Tokyo, Japan.

[Chu et al., 2006] Chu, M., Chen, Y., Zhao, Y., Li, Y., and Soong, F. (2006). A study on how human annotations benefit the TTS voice. In *Proceedings Blizzard Challenge 2006*, pages 999–1000, Pittsburgh, PA, USA.

[Chung and Huckvale, 2001] Chung, H. and Huckvale, M. (2001). Linguistic factors affecting timing in Korean with application to speech synthesis. In *Proceedings Eurospeech*, Aalborg, Denmark.

[Cooper et al., 1983] Cooper, W., Soares, C., Ham, A., and Damon, K. (1983). The influence of inter- and intra-speaker tempo on fundamental frequency and palatalization. *Journal of the Acoustical Society of America (JASA)*, 73:1723–1730.

[Covell et al., 1998] Covell, M., Withgott, M., and Slaney, M. (1998). MACH1 for nonuniform time-scale modification of speech: Theory, technique, and comparisons. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 999–1000, Seattle, WA, USA.

[Crystal and House, 1986] Crystal, T. and House, A. (1986). Characterisation and modeling of speech-segment durations. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2791–2794, Tokyo, Japan.

[Crystal and House, 1988] Crystal, T. and House, A. (1988). The duration of American-English stop consonants: An overview. *Journal of Phonetics*, 16:285–294.

[Crystal and House, 1990] Crystal, T. and House, A. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America (JASA)*, 88:101–112.

[Daniloff and Hammarberg, 1973] Daniloff, R. and Hammarberg, R. (1973). On defining coarticulation. *Journal of Phonetics*, 1:239–248.

[Dankovičová, 1999] Dankovičová, J. (1999). Articulation rate variation within the intonation phrase in Czech and English. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 999–1000, San Francisco, CA, USA.

[Dauer, 1983] Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11:51–62.

[Davidson, 2006] Davidson, L. (2006). Schwa elision in fast speech: Segmental deletion or gestural overlap? *Phonetica*, 63:79–112.

[Delattre, 1966] Delattre, P. (1966). A comparison of syllable length conditioning among languages. *Int. Review of Applied Linguistics*, 4:183–198.

[Dellwo et al., 2006] Dellwo, V., Ferragne, E., and Pellegrino, F. (2006). The perception of intended speech rate in English, French, and German by French speakers. In *Proceedings Speech Prosody*, pages 999–1000, Dresden, Germany.

[Dellwo et al., 2004] Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J., and Wagner, P. (2004). BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate. In *Proceedings Interspeech*, pages 777–780, Barcelona, Spain.

[Dellwo and Wagner, 2003] Dellwo, V. and Wagner, P. (2003). Relations between language rhythm and speech rate. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 471–474, Barcelona, Spain.

[Demenko et al., 2008] Demenko, G., Bachan, J., Möbius, B., Klessa, K., Szymanski, M., and Grocholewski, S. (2008). Development and evaluation of Polish speech corpus for unit selection speech synthesis systems. In *Proceedings Interspeech*, pages 1650–1653, Brisbane, Australia.

[Demenko et al., 2010] Demenko, G., Klessa, K., Szymanski, M., Breuer, S., and Hess, W. (2010). Polish unit selection speech synthesis with BOSS: Extensions and speech corpora. *International Journal of Speech Technology*, 13:85–99.

[Demol et al., 2005] Demol, M., Verhelst, W., Struyve, K., and Verhoeve, P. (2005). Efficient non-uniform time-scaling of speech with WSOLA. In *Proceedings 10th International Conference on Speech Computation*, pages 163–166, Patras, Greece.

[Den Os, 1985] Den Os, E. (1985). Perception of speech rate of Dutch and Italian utterances. *Phonetica*, 42:124–134.

[Dettweiler, 1984] Dettweiler, H. (1984). *Automatisch Sprachsynthese deutscher Wörter mit Hilfe von silbenorientierten Segmenten.* PhD thesis, Technische Universität München.

[Diehl et al., 2004] Diehl, R., Lotto, A., and Holt, L. (2004). Speech perception. http://repository.cmu.edu/psychology/155. Department of Psychology. Paper 155. [Online; accessed 01-March-2010].

[Dietrich et al., 2013] Dietrich, S., Hertrich, I., and Ackermann, H. (2013). Training of ultra-fast speech comprehension induces functional reorganization of the central-visual system in late-blind humans. *Frontiers in Human Neuroscience*, 7.

[Dragon, 2005] Dragon, R. (2005). LAM4HTK. http://www.tnt.uni-hannover.de/d̃ragon/. [Online; accessed 01-March-2010].

[Dressler, 1972] Dressler, W. (1972). Approaches to fast speech rules. *Phonologica*, pages 219–234.

[Dupoux and Green, 1996] Dupoux, E. and Green, K. (1996). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, 23:914–927.

[Eefting and Rietveld, 1989] Eefting, W. and Rietveld, A. (1989). Just noticeable differences of articulation rate at sentence level. *Speech Communication*, 8:355–361.

[Elsendoorn, 1985] Elsendoorn, B. (1985). Acceptability of temporal variations in synthetic speech. A preliminary investigation. *IPO Annual Progress Report, Technische Universiteit Eindhoven*, 20:33–42.

[Engstrand, 1988] Engstrand, O. (1988). Articulatory correlates of stress and speaking rate in Swedish VCV utterances. *Journal of the Acoustical Society of America (JASA)*, 83:1863–1875.

[Engstrand and Krull, 2001] Engstrand, O. and Krull, D. (2001). Segment and syllable reduction: Preliminary observations. *Working Papers Lund, Phonetics Laboratory Lund University*, 49:26–29.

[Fant, 1960] Fant, G. (1960). *Acoustic theory of speech production.* Mouton, The Hague, Netherlands.

[Fant et al., 1991] Fant, G., Kruckenberg, A., and Nord, L. (1991). Some observations on tempo and speaking style in Swedish text reading. In *Proceedings ESCA Workshop on Phonetics and Phonology of Speaking Styles*, pages 231–235, Barcelona, Spain.

[Fant et al., 1992] Fant, G., Kruckenberg, A., and Nord, L. (1992). Prediction of syllable duration, speech rate and tempo. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 667–670, Banff, Alberta, Canada.

[Faust, 1997] Faust, L. (1997). *Variationen von Sprache, ihre Bedeutung für unser Ohr und für die Sprachtechnologie.* PhD thesis, Universität Bonn.

[Fellbaum, 1996] Fellbaum, K. (1996). Einsatz der Sprachsynthese im Behindertenbereich. In *Proceedings Fortschritte der Akustik. DAGA'96*, pages 78–81, Oldenburg.

[Fellbaum and Höpfner,] Fellbaum, K. and Höpfner, D. Anmerkungen zu den Begriffen "Verständlichkeit" und "Verstehbarkeit" bei der Sprachqualitätsmessung. In *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV). Studientexte zur Sprachkommunikation, OPTcrossref = , OPTkey = , pages = 240–247, year = 2014, editor = , volume = , number = , OPTseries = , address = Dresden, Germany, month = , OPTorganization = , publisher = TUD press.*

[Fernandez et al., 2015] Fernandez, R., Rendel, A., Ramabhadran, B., and Hoory, R. (2015). Using deep bidirectional recurrent neual networks for prosodic-target prediction in a unit selection text-to-speech system. In *Proceedings Interspeech 2015*, Dresden, Germany.

[Fosler-Lussier and Morgan, 1998] Fosler-Lussier, E. and Morgan, N. (1998). Effects of speaking rate and word frequency on conversational pronunciations. In *Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 35–40, Rolduc, The Netherlands.

[Fougeron and Jun, 1998] Fougeron, C. and Jun, S. (1998). Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics*, 26:45–69.

[Foulke, 1971] Foulke, E. (1971). The perception of time-compressed speech. In Horton, D. and Jenkins, J., editors, *The perception of language.* Merrill, Columbus, Ohio.

[Foulke and Sticht, 1969] Foulke, E. and Sticht, T. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, 72:50–62.

[Fourakis, 1991] Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America (JASA)*, 90:1816–1827.

[Fowler, 2005] Fowler, C. (2005). Parsing coarticulated speech in perception: Effects of coarticulation resistance. *Journal of Phonetics*, 33:199–213.

[FreedomScientific, 2011] FreedomScientific (2011). Jaws for Windows screen reading software. http://www.freedomscientific.com/products/fs/jaws-product-page.asp. [Online; accessed 01-April-2011].

[Fuchs and Perrier, 2005] Fuchs, S. and Perrier, P. (2005). On the complex nature of speech kinematics. *ZAS Papers in Linguistics*, 42:137–165.

[Fujimura and Lovins, 1978] Fujimura, O. and Lovins, J. (1978). Syllables as concatenative phonetic unit. In Bell, A. and Hooper, J., editors, *Syllables and Segments*. Amsterdam, North Holland.

[Gay, 1968] Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *Journal of the Acoustical Society of America (JASA)*, 44:1570–1573.

[Gay, 1978] Gay, T. (1978). Effect of speaking rate on vowel formant movements. *Journal of the Acoustical Society of America (JASA)*, 63:223–230.

[Gay, 1981] Gay, T. (1981). Mechanisms in the control of speech rate. *Phonetica*, 38:148–158.

[Gers et al., 2002] Gers, F., Schraudolph, N., and Schmidhuber, J. (2002). Learning precise timing with LSTM Recurrent Networks. *Journal of Machine Learning Reserach*, 3:115–143.

[Goldman-Eisler, 1968] Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, New York.

[Gopal, 1990] Gopal, H. (1990). Effects of speaking rate on the behaviour of tense and lax vowel durations. *Journal of Phonetics*, 18:497–518.

[Gottfried et al., 1990] Gottfried, T., Miller, J., and Payton, P. (1990). Effect of speaking rate on the perception of vowels. *Phonetica*, 47:155–172.

[Granström, 1991] Granström, B. (1991). The use of speech synthesis in exploring different speaking styles. *STL-QPSR*, 32:1–10.

[Greenberg, 1996] Greenberg, S. (1996). Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings ESCA Workshop on the Auditory Basis of Speech Perception*, pages 1–8, Keele University, UK.

[Greisbach, 1991] Greisbach, R. (1991). Some aspects of maximally fast reading style. In *Proceedings ESCA Workshop on Phonetics and Phonology of Speaking Styles*, pages 281–285, Barcelona, Spain.

[Greisbach, 1992] Greisbach, R. (1992). Reading aloud at maximal speed. *Speech Communication*, 11:469–473.

[Grice, 1989] Grice, M. (1989). Syntactic structures and lexicon requirements for semantically unpredictable sentences in a number of languages. In *Proceedings of the ESCA workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, Netherlands.

[Grosjean and Deschamps, 1972] Grosjean, F. and Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica*, 26:129–156.

[Grosjean and Deschamps, 1975] Grosjean, F. and Deschamps, A. (1975). Analyse contrastive des variables temporelles de l'anglais et du français: Vitesse de parole et variables composantes, phénomènes d'hésitation. *Phonetica*, 31:144–184.

[Hammerstingl and Breuer, 2003] Hammerstingl, R. and Breuer, S. (2003). Evaluation eines Sprachsynthesesystems nach dem Prinzip der Nonuniform Unit Selection. In *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Karlsruhe, Germany.

[Hammerstingl and Breuer, 2004] Hammerstingl, R. and Breuer, S. (2004). Evaluation eines Sprachsynthesesystems nach dem Prinzip der Non-Uniform Unit Selection. *IKP-Arbeitsberichte Neue Folge*, 10.

[Harris, 1953a] Harris, C. (1953a). A speech synthesizer. *Journal of the Acoustical Society of America (JASA)*, 25:970–975.

[Harris, 1953b] Harris, C. (1953b). A study of the building blocks in speech. *Journal of the Acoustical Society of America (JASA)*, 25:962–969.

[Hasegawa, 1979] Hasegawa, N. (1979). Casual speech vs. fast speech. In Clyne, P., Hanks, W., and Hofbauer, C., editors, *Papers from the Fifteenth Regional Meeting. Chicago, Ill.: Chicago Linguistic Society*, pages 126–137, Chicago, Illinois, USA.

[Hawkins et al., 2000] Hawkins, S., Heid, S., House, J., and Huckvale, M. (2000). Assessment of Naturalness in the ProSynth speech synthesis project. In *Proceedings IEE Colloquium on Speech Synthesis*, London, UK.

[He and Gupta, 2001] He, L. and Gupta, A. (2001). Exploring benefits of non-linear time compression. In *Proceedings of the 9th ACM international conference on Multimedia*, pages 382–391, Ottawa, Canada.

[Hecker and Williams, 1966] Hecker, M. and Williams, C. (1966). Choice of reference conditions for speech preference tests. *Journal of the Acoustical Society of America (JASA)*, 39.

[Heiman et al., 1986] Heiman, G., Leo, R., Leighbody, G., and Bowler, K. (1986). Word intelligibility decrements and the comprehension of time-compressed speech. *Perception & Psychophysics*, 40:407–411.

[Hess, 1992] Hess, W. (1992). Speech synthesis - a solved problem? In Vandewalle, J., Boite, R., Moonen, M., and Oosterlinck, A., editors, *Signal processing VI: Theories and Applications*, pages 37–46. Elsevier, Amsterdam, The Netherlands.

[Hess, 1994] Hess, W. (1994). Sprachsynthese - ein gelöstes Problem? *it - Information Technology*, 36:40–47.

[Hess et al., 1997] Hess, W., Batliner, A., Kießling, A., Kompe, R., Nöth, E., Petzold, A., Reyelt, M., and Strom, V. (1997). Prosodic modules for speech recognition and understanding in VERBMOBIL. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody*, pages 361–382. Springer.

[Hinterleitner et al., 2011] Hinterleitner, F., Moeller, S., Norrenbrock, C., and Heute, U. (2011). Perceptual quality dimensions of text-to-speech systems. In *Proceedings Interspeech 2011*, Florence, Italy.

[Hirata and Tsukada, 2004] Hirata, Y. and Tsukada, K. (2004). The effects of speaking rates and vowel length on formant movements in Japanese. In *Proceedings Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception*, pages 73–85, Somerville, TX, USA.

[Hoequist, 1983] Hoequist, C. (1983). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica*, 40:203–237.

[Hoole et al., 1994] Hoole, P., Mooshammer, C., and Tillmann, H. (1994). Kinematic analysis of vowel production in german. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 53–56, Yokohama, Japan.

[Höpfner, 2007] Höpfner, D. (2007). Untersuchungen zeitskalierter Sprachwiedergabe mit normal sehenden, sehbehinderten und blinden Probanden. In Fellbaum, K., editor, *Elektronische Sprachsignalverarbeitung: Tagungsband der 18. Konferenz, Studientexte zur Sprachkommunikation*, volume 46, pages 235–242, Cottbus/Dresden, Germany. TUD press.

[Höpfner, 2008] Höpfner, D. (2008). Nichtlinearer Zeitskalierungsalgorithmus für gespeicherte natürliche Sprache. In *Proceedings ITG-Fachtagung Sprachkommunikation*, pages 999–1000, Aachen, Germany.

[Huggins, 1972a] Huggins, A. (1972a). Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America (JASA)*, 51:1270–1278.

[Huggins, 1972b] Huggins, A. (1972b). On the perception of temporal phenomena in speech. *Journal of the Acoustical Society of America (JASA)*, 51:1279–1290.

[Huggins, 1979] Huggins, A. (1979). Some effects on intelligibility of inappropriate temporal relations within speech units. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 283–289, Copenhagen, Denmark.

[Hunt and Black, 1996] Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings ICASSP*, pages 373–376.

[Hustad et al., 1998] Hustad, K., Kent, R., and Beukelman, D. (1998). DECTalk and MacinTalk speech synthesizers: Intelligibility differences for three listener groups. *Journal of Speech Language and Hearing Research*, 41:744–752.

[International Phonetic Association, 2005] International Phonetic Association (2005). The international phonetic alphabet. https://www.internationalphoneticassociation.org/content/ipa-chart. [Online; accessed 01-October-2015].

[International Telecommunication Union, 1994] International Telecommunication Union (1994). A method for subjective performance assessment of the quality of speech voice output devices. https://www.itu.int/rec/T-REC-P.85-199406-I/en. [Online; accessed 30-September-2016].

[Jacewicz et al., 2007] Jacewicz, E., Fox, R., and Salmons, J. (2007). Vowel space areas across dialects and gender. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 1465–1468, Saarbrücken, Germany.

[Jannedy et al., 2010] Jannedy, S., Fuchs, S., and Weirich, M. (2010). Articulation beyond the usual: Evaluating the fastes German speaker under laboratory conditions. In Fuchs, S., Hoole, P., and Mooshammer, C.and Zygis, M., editors, *Between the Regular and the Particular in Speech and Language*, pages 205–234. Peter Lang, Frankfurt.

[Janse, 2001] Janse, E. (2001). Comparing word-level intelligibility after linear vs. non-linear time-compression. In *Proceedings Eurospeech*, pages 1407–1410, Aalborg, Denmark.

[Janse, 2002] Janse, E. (2002). Time-compressing natural and synthetic speech. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 1645–1648, Denver, Colorado, USA.

[Janse, 2003a] Janse, E. (2003a). *Production and Perception of Fast Speech*. PhD thesis, Universiteit Utrecht.

[Janse, 2003b] Janse, E. (2003b). Word perception in natural-fast and artificially time-compressed speech. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 3001–3004, Barcelona, Spain.

[Janse, 2004] Janse, E. (2004). Word perception in fast speech. Artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, 42:155–173.

[Janse et al., 2003] Janse, E., Nooteboom, S., and Quené, H. (2003). Word-level intelligibility of time-compressed speech: Prosodic and segmental factors. *Speech Communication*, 41:287–301.

[Janse et al., 2000] Janse, E., Sennema, A., and Slis, A. (2000). Fast speech timing in Dutch: The durational correlates of lexical stress and pitch accent. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 251–254, Beijing, China.

[Janse et al., 2007] Janse, E., van der Werff, M., and Quené, H. (2007). Listening to fast speech: Aging and sentence context. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 681–684, Saarbrücken, Germany.

[Jekosch, 1992] Jekosch, U. (1992). The cluster-identification test. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada.

[Jekosch, 2005] Jekosch, U. (2005). *Voice and speech quality perception. Assessment and evaluation.* Springer Science and Business Media.

[Johnson, 1997] Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In *Talker variability in speech processing*, pages 145–165.

[Jongenburger and van Bezooijen, 1992] Jongenburger, W. and van Bezooijen, R. (1992). Text-to-speech conversion for Dutch: Comprehensibility and acceptability. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 999–1000, Banff, Alberta, Canada.

[Kato et al., 1997] Kato, H., Tsuzaki, M., and Sagisaka, Y. (1997). Measuring temporal compensation in speech perception. In Sagisaka, Y., Campbell, W., and Higuchi, N., editors, *Computing prosody. Computational models for processing spontaneous speech*, pages 251–270. Springer-Verlag, New York, Berlin, Heidelberg.

[Kawai and Toda, 2004] Kawai, H. and Toda, T. (2004). An evaluation of automatic phone segmentation for concatenative speech synthesis. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 999–1000, Montreal, Quebec, Canada.

[Kegel, 1998] Kegel, G. (1998). Störungen der Sprach- und Zeitverarbeitung. Konsequenzen für Diagnose und Therapie. http://www.psycholinguistik.uni-muenchen.de/publ/. [Online; accessed 21-July-2017].

[Keller and Zellner, 1996] Keller, E. and Zellner, B. (1996). A timing model for fast French. *York Papers in Linguistics*, 17:53–75.

[Kent and Read, 1992] Kent, R. and Read, C. (1992). *The Acoustic Analysis of Speech*. Singular Publishing Group.

[Kessinger and Blumstein, 1997] Kessinger, R. and Blumstein, S. (1997). Effects of speaking rate on voice-onset time in Thai, French and English. *Journal of Phonetics*, 25:143–168.

[King et al., 2003] King, S., Black, A., Taylor, P., Caley, R., and Clark, R. (2003). Edinburgh speech tools library. System documentation edition 1.2. http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0. [Online; accessed 01-March-2010].

[Klabbers et al., 2001] Klabbers, E., Stöber, K., Veldhuis, R., Wagner, P., and Breuer, S. (2001). Speech synthesis development made easy: The Bonn Open Synthesis System. In *Proceedings Eurospeech*, pages 999–1000, Aalborg, Denmark.

[Klatt, 1976] Klatt, D. (1976). Linguistic uses of segmental duration in English. *Journal of the Acoustical Society of America (JASA)*, 59:1208–1221.

[Klatt, 1979] Klatt, D. (1979). Synthesis by rule of segmental duration in English sentences. In Lindblom, B. and Öhman, S., editors, *Frontiers of Speech Communication Research. Festschrift for Gunnar Fant*. Academic Press, London, GB.

[Klatt and Klatt, 1990] Klatt, D. and Klatt, L. (1990). Analysis, synthesis and perception of voice quality variations among male and female talkers. *Journal of the Acoustical Society of America (JASA)*, 87:820–856.

[Klessa et al., 2007] Klessa, K., Szymanski, M., Breuer, S., and Demenko, G. (2007). Optimization of Polish segmental duration prediction with CART. In *Proceedings 6th ISCA Speech Synthesis Workshop (SSW-6)*, Bonn, Germany.

[Kohler, 1983] Kohler, K. (1983). Stress-timing and speech rate in German. A production model. *Arbeitsberichte (AIPUK) Institut für Phonetik und digitale Sprachverarbeitung Kiel*, 20:5–53.

[Kohler, 1986] Kohler, K. (1986). Parameters of speech rate perception in German words and sentences: Duration, F0 movement, and F0 level. *Language and Speech*, 29:115–139.

[Kohler, 1988] Kohler, K. (1988). Zeitstrukturierung in der Sprachsynthese. In Lacroix, A., editor, *Proceedings Digitale Sprachverarbeitung, ITG-Fachtagung Sprachkommunikation*, pages 165–170, Bad Nauheim, Germany.

[Kohler, 1990] Kohler, K. (1990). *Segmental reduction in connected speech in German: Phonological facts and phonetic explanations*. Kluwer, Dordrecht.

[Kohler, 1995] Kohler, K. (1995). Articulatory reduction in different speaking styles. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 12–19, Stockholm, Sweden.

[Kohler et al., 1981] Kohler, K., Schäfer, K., Thon, W., and Timmermann, G. (1981). Sprechgeschwindigkeit in Produktion und Perzeption. *Arbeitsberichte (AIPUK) Institut für Phonetik und digitale Sprachverarbeitung Kiel*, 16:137–179.

[Kominek et al., 2003] Kominek, J., Bennett, C., and Black, A. (2003). Evaluating and correcting phoneme segmentation for unit selection synthesis. In *Proceedings Eurospeech*, pages 999–1000, Geneva, Switzerland.

[Koopmans-Van Beinum, 1990] Koopmans-Van Beinum, J. (1990). Spectro-temporal reduction and expansion in spontaneous speech and read text: The role of focus words. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Kobe, Japan.

[Koreman, 2003] Koreman, J. (2003). The perception of articulation rate. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 999–1000, Barcelona, Spain.

[Koreman, 2006] Koreman, J. (2006). The role of articulation rate in distinguishing fast and slow speakers. In *Proceedings Speech Prosody*, Dresden, Germany.

[Kraft and Portele, 1995] Kraft, V. and Portele, T. (1995). Quality evaluation of five German speech synthesis systems. *Acta Acustica*, 3:351–365.

[Krause and Braida, 2002] Krause, J. and Braida, L. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America (JASA)*, 112:2165–2172.

[Krishna and Murthy, 2004] Krishna, N. and Murthy, H. (2004). Duration modeling of Indian languages Hindi and Telugu. In *Proceedings 5th ISCA Speech Synthesis Workshop (SSW-5)*, Pittsburgh, PA.

[Kuehn and Moll, 1976] Kuehn, D. and Moll, K. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics*, 4:303–320.

[Kuhl, 1991] Kuhl, P. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Attention, Perception, & Psychophysics*, 50:93–107.

[Kuwabara, 1997] Kuwabara, H. (1997). Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. In *Proceedings Eurospeech*, Rhodes, Greece.

[Laver, 1994] Laver, J. (1994). *Principles of phonetics*. Cambridge University Press, Cambridge, MA, USA.

[Lebeter and Saunders, 2010] Lebeter, J. and Saunders, S. (2010). The effects of time compression on the comprehension of natural and synthetic speech. *Working Papers of the Linguistics Circle*, 20:63–81.

[Lehiste, 1972] Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America (JASA)*, 51:2018–2024.

[Lehiste, 1977] Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5:253–263.

[Lehiste, 1994] Lehiste, I. (1994). *Suprasegmentals*. MIT Press, Cambridge, MA, USA.

[Lehiste and Peterson, 1961] Lehiste, I. and Peterson, G. (1961). Transitions, glides, and diphthongs. *Journal of the Acoustical Society of America (JASA)*, 33:268–277.

[Lemetty, 1999] Lemetty, S. (1999). Review of speech synthesis technology. *Helsinki University of Technology*, 320:79–90.

[Lewandowski, 2011] Lewandowski, N. (2011). *Talent in nonnative phonetic convergence*. PhD thesis, Stuttgart University, Institut für Maschinelle Sprachverarbeitung.

[Liberman, 1981] Liberman, A. (1981). On finding that speech is special. *Status Report on Speech Research*, SR-6768:107–143.

[Liberman, 1996] Liberman, A. (1996). *Speech: A special code*. MIT press.

[Lindblom, 1963] Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America (JASA)*, 35:1773–1781.

[Lindblom, 1983] Lindblom, B. (1983). Economy of speech gestures. In Lindblom, B., editor, *The production of speech*, pages 217–245. Springer.

[Lindblom, 1990] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H-Theory. In Hardcastle, W. and Marchal, A., editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer, Dordrecht.

[Lindblom, 1996] Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America (JASA)*, 99:1683–1692.

[Liu et al., 2008] Liu, C., Hsu, H., and Lee, W. (2008). Compression artifacts in perceptual audio coding. In *IEEE Transactions on Audio, Speech, and Language Processing*, volume 16(4), pages 681–695.

[Liu and Zeng, 2006] Liu, S. and Zeng, F. (2006). Temporal properties in clear speech perception. *Journal of the Acoustical Society of America (JASA)*, 120:424–432.

[Loizou, 2011] Loizou, P. (2011). Speech quality assessment. In Lin, W. e. a., editor, *Multimedia Analysis, Processing and Communications*, pages 623–654. Springer Verlag, Berlin, Heidelberg.

[Luce and Pisoni, 1983] Luce, P. and Pisoni, D. (1983). Capacity-demanding encoding of synthetic speech in serial-ordered recall. *Research on Speech Perception Progress Report*, 9.

[Maniwa et al., 2009] Maniwa, K., Jongman, A., and Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *Journal of the Acoustical Society of America (JASA)*, 125:3962–3973.

[Marslen-Wilson and Tyler, 1980] Marslen-Wilson, W. and Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition*, 8:1–71.

[Martinez et al., 1997] Martinez, F., Tapias, D., Alvarez, J., and Leon, P. (1997). Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *Proceedings Eurospeech*, pages 999–1000, Rhodes, Greece.

[McCarthy et al., 2013] McCarthy, T., Pal, J., and Cutrell, E. (2013). The ”voice” has it: Screen reader adoption and switching behavior among vision impaired persons in India. *Assist Technology: The Official Journal of RESNA*, 25(4):222–229.

[Mefferd and Green, 2010] Mefferd, A. and Green, J. (2010). Articulatory-to-acoustic relations in response to speaking rate and loudness manipulations. *Journal of Speech, Language, and Hearing Research*, 53:1206–1219.

[Merritt et al., 2016] Merritt, T., Clark, R., Wu, Z., Yamagishi, J., and King, S. (2016). Deep neural network-guided unit selection synthesis. In *Proceedings ICASSP*.

[Meyer et al., 1995] Meyer, H., Portele, T., and Heuft, B. (1995). Ein Silbendauermodell für ein Sprachsynthesesystem. In *Fortschritte der Akustik*, pages 987–990, Saarbrücken, Germany.

[Miller and Baer, 1983] Miller, J. and Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America (JASA)*, 73:1751–1755.

[Miller et al., 1986] Miller, J., Green, K., and Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, 43:106–115.

[Miller et al., 1984] Miller, J., Grosjean, F., and Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, 41:215–225.

[Miller and Liberman, 1979] Miller, J. and Liberman, A. (1979). Some effects of later-occurring information on the perception of stop consonants and semi-vowels. *Perception & Psychophysics*, 25:457–465.

[Miller et al., 1997] Miller, J., O'Rourke, T., and Volaitis, L. (1997). Internal structure of phonetic categories: Effects of speaking rate. *Phonetica*, 54:121–137.

[Miller and Volaitis, 1989] Miller, J. and Volaitis, L. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, 46:505–512.

[Mixdorff et al., 2005] Mixdorff, H., Pfitzinger, H., and Grauwinkel, K. (2005). Towards objective measures for comparing speaking styles. In *Proceedings SPECOM 2005*, pages 131–134.

[Möbius, 1995] Möbius, B. (1995). Components of a quantitative model of German intonation. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 108–115, Philadelphia, PA, USA.

[Möbius, 2000] Möbius, B. (2000). Corpus-based speech synthesis: Methods and challenges. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Universität Stuttgart*, 6:87–116.

[Möbius, 2003] Möbius, B. (2003). Rare events and closed domains: Two delicate concepts in speech synthesis. *International Journal of Speech Technology*, 6:57–71.

[Möbius and van Santen, 1996] Möbius, B. and van Santen, J. (1996). Modeling segmental duration in German text-to-speech synthesis. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 2395–2398, Philadelphia, PA, USA.

[Moers, 2011] Moers, D. (2011). Schnell gesprochene Sprache als Einheiten-Auswahl-Inventar in der Unit-Selection-Sprachsynthese. In *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Aachen, Germany.

[Moers and Wagner, 2008] Moers, D. and Wagner, P. (2008). Evaluation eines Sprechers für schnell gesprochene Sprache in der Unit-Selection basierten Sprachsynthese. In *Proceedings ITG-Fachtagung Sprachkommunikation*, Aachen, Germany.

[Moers and Wagner, 2009] Moers, D. and Wagner, P. (2009). Assessing a speaker for fast speech in unit selection speech synthesis. In *Proceedings Interspeech 2009*, Brighton, UK.

[Moers et al., 2007] Moers, D., Wagner, P., and Breuer, S. (2007). Assessing the adequate treatment of fast speech in unit selection speech synthesis systems for the visually impaired. In *Proceedings 6th ISCA Workshop on Speech Synthesis (SSW-6)*, Bonn, Germany.

[Moers et al., 2010a] Moers, D., Wagner, P., and Möbius, B. (2010a). Erzeugung schnell gesprochener Sprache in der Unit Selection Sprachsynthese. In *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Berlin, Germany.

[Moers et al., 2010b] Moers, D., Wagner, P., Möbius, B., and Jauk, I. (2010b). Synthesizing fast speech by implementing multi-phone units in unit selection speech synthesis. In *Proceedings 7th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW-7)*, Kyoto, Japan.

[Moers et al., 2010c] Moers, D., Wagner, P., Möbius, B., Müllers, F., and Jauk, I. (2010c). Integrating a fast speech corpus in unit selection speech synthesis: Experiments on perception, segmentation and duration prediction. In *Proceedings Speech Prosody 2010*, Chicago, IL.

[Moers et al., 2010d] Moers, D., Wagner, P., Möbius, B., Müllers, F., and Jauk, I. (2010d). Schnell gesprochene Sprache in der Unit-Selection-Sprachsynthese: Untersuchungen zu Korpuserstellung und -aufbereitung. In *Proceedings ITG-Fachtagung Sprachkommunikation*, Bochum, Germany.

[Möhler, 1998] Möhler, G. (1998). *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese.* PhD thesis, Universität Stuttgart.

[Mok, 2007] Mok, P. (2007). *Influences on vowel-to-vowel coarticulation.* PhD thesis, University of Cambridge.

[Möller, 2000] Möller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications.* Springer Science & Business Media.

[Möller et al., 2010] Möller, S., Hinterleitner, F., Falk, T., and Polzehl, T. (2010). Comparison of approaches for intrumentally predicting the quality of text-to-speech systems. In *Proceedings of Interspeech 2010*, Makuhari, Japan.

[Monaghan, 2001] Monaghan, A. (2001). An auditory analysis of the prosody of fast and slow speech styles in English, Dutch and German. In Keller, E., Bailly, G., Monaghan, A., et al., editors, *Improvements in Speech Synthesis*, pages 204–217. Kluwer, Chichester.

[Moon and Lindblom, 1994] Moon, S. and Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America (JASA)*, 96:40–55.

[Moos et al., 2008] Moos, A., Hertrich, I., Dietrich, S., Trouvain, J., and Ackermann, H. (2008). Perception of ultra fast speech by a blind listener: Does he use his visual system? In *Proceedings of International Seminar on Speech Production (ISSP)*, pages 297–300, Strasbourg, France.

[Moos and Trouvain, 2007] Moos, A. and Trouvain, J. (2007). Comprehension of ultra-fast speech: Blind vs. "normally hearing" persons. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 677–684, Saarbrücken, Germany.

[Moos and Trouvain, 2008] Moos, A. and Trouvain, J. (2008). Einzelfallstudie zu Grenzen der Verständlichkeit ultra-schneller Sprachsynthese. In *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, pages 207–214, Frankfurt/Main, Germany.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467.

[Neppert and Petursson, 1986] Neppert, J. and Petursson, M. (1986). *Elemente einer akustischen Phonetik*. Helmut Buske Verlag, Hamburg.

[Nishimoto et al., 2006] Nishimoto, T., Sako, S., Sagayama, S., Ohshima, K., Oda, K., and Watanabe, T. (2006). Effect of learning on listening to ultra-fast synthesized speech. In *Proceedings 28th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society*, pages 5691–5694, New York City, NY, USA.

[Nooteboom, 1972] Nooteboom, S. (1972). *Production and Perception of vowel duration: A study of durational properties of vowels in Dutch*. PhD thesis, Rijksuniversiteit Utrecht.

[Nooteboom and Eefting, 1994] Nooteboom, S. and Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica*, 51:92–98.

[Nygaard et al., 1994] Nygaard, L., Sommers, M., and Pisoni, D. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5:42–46.

[Öhman, 1965] Öhman, S. (1965). On the coordination of articulatory and phonatory activity in the production of Swedish tonal accents. *STL Progress and Status Report No. 2*, pages 14–19.

[Öhman, 1967] Öhman, S. (1967). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustic Society of America (JASA)*, 39:151–168.

[Ohno and Fujisaki, 1995] Ohno, S. and Fujisaki, H. (1995). A method for quantitative analysis of the local speech rate. In *Proceedings Eurospeech*, pages 421–424, Madrid, Spain.

[Ohno et al., 1997] Ohno, S., Fujisaki, H., and Taguchi, H. (1997). A method for quantitative analysis of the local speech rate using an inventory of reference units. In *Proceedings Eurospeech*, pages 461–464, Rhodes, Greece.

[Okadome et al., 1999] Okadome, T., Kaburagi, T., and Honda, M. (1999). Relations between utterance speed and articulatory movements. In *Proceedings Eurospeech*, Budapest, Hungary.

[Olive, 1990] Olive, J. (1990). A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. In *Proceedings ESCA Workshop on Speech Synthesis*, pages 25–30, Autrans, France.

[Orr et al., 1965] Orr, D., Friedman, H., and Williams, J. (1965). Trainability of listening comprehension of speeded discourse. *Journal of Educational Psychology*, 56:148–156.

[Osser and Peng, 1964] Osser, H. and Peng, F. (1964). A cross cultural study of speech rate. *Language and Speech*, 7:120–125.

[Ostry and Munhall, 1985] Ostry, D. and Munhall, K. (1985). Control of rate and duration of speech movements. *Journal of the Acoustic Society of America (JASA)*, 77:640–648.

[Pallier et al., 1998] Pallier, C., Sebastian-Gallés, N., Dupoux, E., Christophe, A., and Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition*, 26:844–851.

[Papadopoulos et al., 2010] Papadopoulos, K., Katemidou, E., Koutsoklenis, A., and Mouratidou, E. (2010). Differences amongst sighted individuals and individuals with visual impairments in word intelligibility presented via synthetic and natural speech. *Augmentative and Alternative Communication*, 26:278–288.

[Papadopoulos and Koustriava, 2015] Papadopoulos, K. and Koustriava, E. (2015). Comprehension of synthetic and natural speech: Differences among sighted and visually impaired young adults. In *Proceedings ICEAPVI*, pages 147–151, Athens, Greece.

[Pasdeloup et al., 2008] Pasdeloup, V., Espesser, R., Piotrowski, D., and Faraj, M. (2008). Form and substance relationship in rhythmic structuring. A morphodynamic analysis of rate sensitivity at the infra-syllabic level. In *Proceedings Speech Prosody*, pages 367–370, Campinas, Brazil.

[Peterson and Lehiste, 1960] Peterson, G. and Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America (JASA)*, 32:693–703.

[Peterson et al., 1958] Peterson, G., Wang, W., and Sievertsen, E. (1958). Segmentation techniques in speech synthesis. *Journal of the Acoustical Society of America (JASA)*, 30:739–742.

[Pfitzinger, 1996] Pfitzinger, H. (1996). Two approaches to speech rate estimation. In *Proceedings SST*, pages 421–426, Adelaide, Australia.

[Pfitzinger, 1998] Pfitzinger, H. (1998). Local speech rate as a combination of syllable and phone rate. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 1087–1090, Sydney, Australia.

[Pfitzinger, 1999] Pfitzinger, H. (1999). Local speech rate perception in German speech. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 893–896, San Francisco, CA, USA.

[Pfitzinger, 2001] Pfitzinger, H. (2001). Phonetische Analyse der Sprechgeschwindigkeit. *Forschungsberichte Institut für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, 38:117–264.

[Pfitzinger et al., 1996] Pfitzinger, H., Burger, S., and Heid, S. (1996). Syllable detection in read and spontaneous speech. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 1261–1264, Philadelphia, Pennsylvania, USA.

[Pickett and Pollack, 1963] Pickett, J. and Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, 6:151–164.

[Pierrehumbert, 2000] Pierrehumbert, J. (2000). Exemplar dynamics: Word frequency, lenition, and contrast. Frequency effects and emergent grammar. John Benjamins, Amsterdam, The Netherlands.

[Pike, 1945] Pike, K. (1945). *The Intonation of American English.* University of Michigan Press, Ann Arbor, Michigan.

[Pimsleur et al., 1977] Pimsleur, P., Hancock, C., and Furey, P. (1977). Speech rate and listening comprehension. Viewpoints on English as a second language. Regent, New York, NY.

[Pisoni, 1981] Pisoni, D. (1981). Speeded classification of natural and synthetic speech in a lexical decision task. *Journal of the Acoustical Society of America (JASA)*, 70.

[Pisoni, 1993] Pisoni, D. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech communication*, 13:109–125.

[Pisoni and Hunnicut, 1980] Pisoni, D. and Hunnicut, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 572–575, New York, USA.

[Pollet et al., 2017] Pollet, V., Zovato, E., Sufian, I., and Batzu, P. (2017). Unit selection with hierarchical cascaded long short term memory bidirectional recurrent neural nets. In *Proceedings Interspeech*, pages 3966–3970, Stockholm, Sweden.

[Pols, 1989] Pols, L. (1989). Improving synthetic speech quality by systematic evaluation. In *Proceedings ESCA Tutorial and Research Workshop on Speech InputOutput Assessment and Speech Databases*, Noordwijkerhout, The Netherlands.

[Pols, 1992] Pols, L. (1992). Multi-lingual synthesis evaluation methods. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Banff, Alberta, Canada.

[Pols, 1999] Pols, L. (1999). Flexible, robust, and efficient human speech processing versus present-day speech technology. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 9–16, San Francisco, CA.

[Pols and Jekosch, 1994] Pols, L. and Jekosch, U. (1994). A structured way of looking at the performance of text-to-speech systems. In *Proceedings 2nd ISCA Speech Synthesis Workshop (SSW-2)*, pages 203–206, NY, USA.

[Pols and van Son, 1993] Pols, L. and van Son, R. (1993). Acoustics and perception of dynamic vowel segments. *Speech Communication*, 13:135–147.

[Port, 1981] Port, R. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America (JASA)*, 69(1):262–274.

[Port and Dalby, 1982] Port, R. and Dalby, J. (1982). C/V ratio as a cue for voicing in English. *Perception and Psychophysics*, 16:257–282.

[Portele, 1996] Portele, T. (1996). *Ein phonetisch-akustisch motiviertes Inventar zur Sprachsynthese deutscher Äußerungen.* PhD thesis, Universität Bonn, Tübingen: Niemeyer.

[Portele, 1997] Portele, T. (1997). Reduktionen in der einheitenbasierten Sprachsynthese. In *Proceedings Fortschritte der Akustik DAGA*, pages 386–387, Kiel, Germany.

[Portele et al., 1994] Portele, T., Heuft, B., Höfer, F., Meyer, H., and Hess, W. (1994). A new high quality speech synthesis system for German. In *Proceedings CRIM/FORWISS Workshop: Progress and Prospects of Speech Research and Technology*, pages 274–277, Munich, Germany.

[Portele and Krämer, 1996] Portele, T. and Krämer, J. (1996). Adapting a TTS system to a reading machine for the blind. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 999–1000, Philadelphia, PA, USA.

[Prieto and Torreira, 2007] Prieto, P. and Torreira, F. (2007). The segmental anchoring hypothesis revisited. Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics*, 35:473–500.

[Quené, 1996] Quené, H. (1996). Apollo-spraaksynthese. http://www.let.uu.nl/H̃ugo.Quene/personal/demos/apollo.html. [Online; accessed 02-July-2015].

[Quené, 2007] Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, 35:353–362.

[RDevelopmentCoreTeam, 2011] RDevelopmentCoreTeam (2011). R: A language and environment for statistical computing, R foundation for statistical computing. http://www.R-project.org. [Online; accessed 01-July-2011].

[Reynolds et al., 2002] Reynolds, M., Isaacs-Duvall, C., and Haddox, M. (2002). A comparison of learning curves in natural and synthesized speech comprehension. *Journal of Speech, Language and Hearing Research*, 45:802–810.

[Riley, 1990] Riley, M. (1990). Tree-based modelling for speech synthesis. In *Proceedings ESCA Workshop on Speech Synthesis*, pages 229–232, Autrans, France.

[Roach, 1998] Roach, P. (1998). Some languages are spoken more quickly than others. In Bauer, L. and Trudgill, P., editors, *Language Myths*, pages 150–158. Penguin.

[Roach et al., 1992] Roach, P., Sergeant, P., and Miller, D. (1992). Syllabic consonants at different speaking rates: A problem for automatic speech recognition. *Speech Communication*, 11:475–479.

[Roodenrys et al., 2002] Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., and Nimmo, L. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6):1019–1034.

[Ruske and Schotola, 1978] Ruske, G. and Schotola, T. (1978). An approach to speech recognition using syllabic decision units. In *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 722–725, Tulsa, OA, USA.

[Sagisaka, 1988] Sagisaka, Y. (1988). Speech synthesis by rule using an optimal selection of nonuniform synthesis units. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 679–682, New York, NY.

[Samlowski et al., 2013] Samlowski, B., Wagner, P., and Möbius, B. (2013). Effects of lexical class and lemma frequency on German homographs. In *Proceedings of Interspeech 2013*, pages 597–601, Lyon, France.

[Sanderman and Collier, 1997] Sanderman, A. and Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech*, 40:391–409.

[Schiel et al., 2006] Schiel, F., Draxler, C., Ellbogen, T., Jänsch, K., and Schmidt, S. (2006). Die BITS Sprachsynthesekorpora - Diphon- und Unit Selection-Synthesekorpora für das Deutsche. In *Proceedings KONVENS*, pages 121–124, Konstanz, Germany.

[Schindler, 1975] Schindler, F. (1975). Faktoren phonetischer Performanz. Instrumentalphonetische Versuche zur akustischen Bestimmung des Ausprägungsgrades von Eigenschaften des lautsprachlichen Signals. *Zeitschrift für Dialektologie und Linguistik. Beihefte. Neue Folge Nr. 14 der Zeitschrift für Mundartforschung.*, 14.

[Schlink, 1994] Schlink, B. (1994). *Selbs Betrug.* detebe.

[Scholtz, 2004] Scholtz, J. (2004). Usability evaluation. *National Institute of Standards and Technology*, 1.

[Schwab et al., 1985] Schwab, E., Nusbaum, H., and Pisoni, D. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27:395–408.

[Schweitzer et al., 2004] Schweitzer, A., Braunschweiler, N., Dogil, G., and Möbius, B. (2004). Assessing the acceptability of the SmartKom speech synthesis voices. In *Proceedings 5th ISCA Speech Synthesis Workshop (SSW-5)*, pages 1–6, Pittsburgh, PA, USA.

[Shen et al., 2017] Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., Saurous, R., Agiomyrgiannakis, Y., and Wu, Y. (2017). Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. *arXiv preprint arXiv*, 1712.05884.

[Sityaev et al., 2006] Sityaev, D., Knill, K., and Burrows, T. (2006). Comparison of the ITU-T P.85 standard to other methods for the evaluation of text-to-speech systems. In *International Conference on Spoken Language Processing (ICSLP)*, pages 1077–1080, Pittsburgh, Pennsylvania.

[Slowiaczek and Nusbaum, 1985] Slowiaczek, L. and Nusbaum, H. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*, 27:701–752.

[Slowiaczek and Pisoni, 1982] Slowiaczek, L. and Pisoni, D. (1982). Effects of practice on speeded classification of natural and synthetic speech. *Research on Speech Perception Progress Report*, 7:255–262.

[Smith et al., 1975] Smith, B., Brown, B., Strong, W., and Rencher, A. (1975). Effects of speech rate on personality perception. *Language and Speech*, 18:145–152.

[Sonntag, 1999] Sonntag, G. (1999). *Evaluation von Prosodie.* PhD thesis, Universität Bonn, Aachen: Shaker Verlag.

[Sotscheck, 1982] Sotscheck, J. (1982). Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte. *Der Fernmeldeingenieur*, 36:1–84.

[Sproat and Olive, 1995] Sproat, R. and Olive, J. (1995). Text-to-speech synthesis. *AT&T technical journal*, 74:35–44.

[Stent et al., 2011] Stent, A., Syrdal, A., and Mishra, T. (2011). On the intelligibility of fast synthesized speech for individuals with early-onset blindness. In *Proceedings ACM ASSETS'11*, Dundee, Scotland, UK.

[Stöber, 2002] Stöber, K. (2002). *Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese.* PhD thesis, Universität Bonn, FrankfurtM.: P. Lang.

[Stöber et al., 1999] Stöber, K., Portele, T., Wagner, P., and Hess, W. (1999). Synthesis by word concatenation. In *Proceedings Eurospeech*, pages 619–622, Budapest, Hungary.

[Stöber et al., 2000] Stöber, K., Wagner, P., Helbig, J., Köster, S., Stall, D., Thomae, M., Blauert, J., Hess, W., Hoffmann, R., and Mangold, H. (2000). Speech synthesis using multilevel selection and concatenation of units from large speech corpora. In Wahlster, W., editor, *Verbmobil: Foundations of speech-to-speech translation*. Springer, Berlin, Germany.

[Summerfield, 1981] Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 77:1074–1095.

[Syrdal et al., 2012] Syrdal, A., Bunnell, H., Hertz, S., Mishra, T., Spiegel, M., Bickley, C., Rekart, D., and Makashay, M. (2012). Text-to-speech intelligibility across speech rates. In *Proceedings Interspeech*, pages 623–626, Portland, OR.

[Syrdal et al., 1997] Syrdal, A., Conkie, A., Stylianou, Y., Schroeter, J., Garrison, L., and Dutton, D. (1997). Voice selection for speech synthesis. *Journal of the Acoustical Society of America (JASA)*, 102(5):3181–3191.

[Syrdal et al., 1998] Syrdal, A., Stylianou, Y., Garrison, L., Conkie, A., and Schroeter, J. (1998). TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis. In *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 273–276, Seattle, WA.

[Tauroza and Allison, 1990] Tauroza, S. and Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, 11:90–105.

[Traunmüller, 2000] Traunmüller, H. (2000). Evidence for demodulation in speech perception. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

[Trouvain, 2002a] Trouvain, J. (2002a). Tempo control in speech synthesis by prosodic phrasing. In *Proceedings Konvens*, Saarbrücken, Germany.

[Trouvain, 2002b] Trouvain, J. (2002b). Temposteuerung in der Sprachsynthese durch prosodische Phrasierung. In *Proceedings Konvens*, Saarbrücken, Germany.

[Trouvain, 2004] Trouvain, J. (2004). *Tempo Variation in Speech Production. Implications for Speech Synthesis.* PhD thesis, Universität des Saarlandes.

[Trouvain, 2006] Trouvain, J. (2006). Subjektive Verständlichkeit von Computerstimmen bei verschiedenen Geschwindigkeiten. Eine Pilotstudie mit zwei Benutzergruppen. Personal communication.

[Trouvain, 2007] Trouvain, J. (2007). Comprehension of synthetic speech. *Saarland Working Papers in Linguistics (SWPL)*, 1:5–13.

[Trouvain and Grice, 1999] Trouvain, J. and Grice, M. (1999). The effect of tempo on prosodic structure. In *Proceedings International Congress of Phonetic Sciences (ICPhS)*, pages 1067–1070, San Francisco, CA, USA.

[Trouvain et al., 2001] Trouvain, J., Koreman, J., Erriquez, A., and Braun, B. (2001). Articulation rate measures and their relations to phone classification of spontaneous and read German speech. In *Proceedings ISCA Workshop on Adaptation Methods for Speech Recognition*, pages 155–158, Sophia Antipolis, France.

[Trouvain et al., 2008] Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., and Barry, W. (2008). Modelling personality features by changing prosody in synthetic speech. In *Proceedings Speech Prosody*, Dresden, Germany.

[Tsao and Weismer, 1997] Tsao, Y.-C. and Weismer, G. (1997). Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language and Hearing research*, 40:858–866.

[Tucker and Whittaker, 2006] Tucker, S. and Whittaker, S. (2006). Time is of the essence. An evaluation of temporal compression algorithms. In ACM, editor, *Proceedings SIGCHI conference on Human Factors in computing systems*, pages 329–338, Montréal, Québec, Canada.

[Vaane, 1982] Vaane, E. (1982). Subjective estimation of speech rate. *Phonetica*, 39:136–149.

[van Bergem, 1993] van Bergem, D. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, 12:1–23.

[van Bergem, 1995] van Bergem, D. (1995). Experimental evidence for a comprehensive theory of vowel reduction. In *Proceedings Eurospeech*, pages 999–1000, Madrid, Spain.

[van den Oord et al., 2016] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv*, 1609.03499.

[van Heuven and van Bezooijen, 1995] van Heuven, V. and van Bezooijen, R. (1995). Quality evaluation of synthesized speech. In Kleijn, W. and Paliwal, K., editors, *Speech Coding and Synlhesis*, pages 707–738. Elsevier.

[van Santen, 1992] van Santen, J. (1992). Deriving text-to-speech durations from natural speech. In Bailly, G. and Benoit, C., editors, *Talking Machines: Theories, Models & Application*. Elsevier.

[van Santen, 1993] van Santen, J. (1993). Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language*, 7:49–100.

[van Santen, 1994] van Santen, J. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128.

[van Santen, 1997] van Santen, J. (1997). Segmental duration and speech timing. In Sagisaka, Y., Campbell, W., and Higuchi, N., editors, *Computing prosody. Computational models for processing spontaneous speech*, pages 225–249. Springer-Verlag, New York, Berlin, Heidelberg.

[van Santen, 1998] van Santen, J. (1998). Quantitative modeling of segmental duration. In Association for Computational Linguistics (ACL), editor, *Proceedings Workshop on Human Language Technology*, pages 323–328, Stroudsburg, PA, USA.

[van Santen and Buchsbaum, 1997] van Santen, J. and Buchsbaum, A. (1997). Methods for optimal text selection. In *Proceedings of Eurospeech*, pages 553–556, Rhodes, Greece.

[van Son and Pols, 1989] van Son, R. and Pols, L. (1989). Comparing formant movements in fast and normal rate speech. In *Proceedings Eurospeech*, pages 2665–2668, Paris, France.

[van Son and Pols, 1995] van Son, R. and Pols, L. (1995). What does consonant reduction look like, if it exists? In *Proceedings Eurospeech*, pages 1909–1912, Madrid, Spain.

[van Son and Pols, 1996] van Son, R. and Pols, L. (1996). An acoustic profile of consonant reduction. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 1529–1532, Philadelphia, USA.

[van Son and Pols, 1999] van Son, R. and Pols, L. (1999). An acoustic description of consonant reduction. *Speech Communication*, 28:125–140.

[Venkatagiri, 1994] Venkatagiri, H. (1994). Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication*, 10:96–104.

[Venkatagiri, 2003] Venkatagiri, H. (2003). Segmental intelligibility of four currently used text-to-speech synthesis methods. *Journal of the Acoustical Society of America (JASA)*, 113:2095–2104.

[Verbrugge and Shankweiler, 1977] Verbrugge, R. and Shankweiler, D. (1977). Prosodic information for vowel identity. *Journal of the Acoustical Society of America (JASA)*, 61:39.

[Verbrugge et al., 1976] Verbrugge, R., Strange, W., Shankweiler, D., and Edman, T. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America (JASA)*, 60:198–212.

[Voiers, 1977] Voiers, W. (1977). Diagnostic evaluation of speech intelligibility. *Benchmark papers in acoustics*, 11.

[Volaitis and Miller, 1992] Volaitis, L. and Miller, J. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America (JASA)*, 92:723–735.

[von Essen, 1949] von Essen, O. (1949). Das Sprechtempo als Ausdruck psychischen Geschehens. *Zeitschrift für Phonetik*, 4:317–340.

[Voor and Miller, 1965] Voor, J. and Miller, J. (1965). The effect of practice upon the comprehension of time-compressed speech. *Speech Monographs*, 32:452–455.

[Wade et al., 2010] Wade, T., Dogil, G., Schütze, H., Walsh, M., and Möbius, B. (2010). Syllable frequency effects in a context-sensitive segment production model. *Journal of Phonetics*, 38:227–239.

[Wade and Möbius, 2007] Wade, T. and Möbius, B. (2007). Speaking rate effects in a landmark-based phonetic exemplar model. In *Proceedings Interspeech*, pages 402–405, Antwerpen, Belgium.

[Wagner, 2005] Wagner, P. (2005). Great Expectations - Introspective vs. Perceptual Prominence Ratings and their Acoustic Correlates. In *Proceedings Interspeech*, pages 2381–2384, Lisbon, Portugal.

[Wagner, 2013] Wagner, P. (2013). (What is) the contribution of phonetics to contemporary speech synthesis(?). In *Systemtheorie. Signalverarbeitung. Sprachtechnologie. Rüdiger Hoffmann zum 65. Geburtstag. Studientexte zur Sprachkommunikation*, volume 68. TUD press, Dresden, Germany.

[Wagner et al., 1999] Wagner, P., Haas, F., Stöber, K., and Helbig, J. (1999). Multilinguale korpusbasierte Sprachsynthese auf der Basis domänenspezifischen Ausgangsmaterials. In *Tagungsband Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, volume 16, pages 152–159, Görlitz, Germany.

[Wan et al., 2017] Wan, V., Agiomyrgiannakis, Y., Silen, H., and Vit, J. (2017). Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence LSTM-based autoencoders. In *Proceedings Interspeech*, pages 1143–1147, Stockholm, Sweden.

[Wang et al., 2000] Wang, C., Fujisaki, H., Tomana, R., and Ohno, S. (2000). Analysis of fundamental frequency contours of standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

[Wayland et al., 1994] Wayland, S., Miller, J., and Volaitis, L. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America (JASA)*, 95:2694–2701.

[Weismer and Berry, 2003] Weismer, G. and Berry, J. (2003). Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *Journal of the Acoustical Society of America (JASA)*, 113:3362–3378.

[Weiss, 2008] Weiss, B. (2008). *Sprechtempoabhängige Aussprachevariationen.* PhD thesis, Humboldt-Universität Berlin.

[Wells, 1997] Wells, J. (1997). SAMPA computer readable phonetic alphabet. http://www.phon.ucl.ac.uk/home/sampa/. [Online; accessed 01-March-2010].

[Whalen, 1990] Whalen, D. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18:3–35.

[Widera, 2000] Widera, C. (2000). Strategies of vowel reduction - A speaker-dependent phenomenon. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, pages 999–1000, Beijing, China.

[Widera, 2003] Widera, C. (2003). *Zur Reduktion von Vokalen: Eine experimentalphonetische Untersuchung.* PhD thesis, Universität Bonn, Göttingen: Cuvillier.

[Widera and Portele, 1999] Widera, C. and Portele, T. (1999). Levels of reduction for German tense vowels. In *Proceedings Eurospeech*, pages 999–1000, Budapest, Hungary.

[Windmann et al., 2013] Windmann, A., Simko, J., Wrede, B., and Wagner, P. (2013). Modeling durational incompressibility. In *Proceedings Interspeech.*

[Wingfield et al., 1984] Wingfield, A., Lombardi, L., and Sokol, S. (1984). Prosodic features and the intelligibility of accelerated speech: Syntactic vs. periodic segmentation. *Journal of Speech and Hearing Research*, 27:128–134.

[Wingfield and Nolan, 1980] Wingfield, A. and Nolan, K. (1980). Spontaneous segmentation in normal and time-compressed speech. *Perception and Psychophysics*, 28:97–102.

[Winters and Pisoni, 2004] Winters, S. and Pisoni, D. (2004). Perception and comprehension of synthetic speech. *Progress Report No. 26*, 26:95–138.

[Wolters et al., 2010] Wolters, M., Isaac, K., and Renals, S. (2010). Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proceedings 7th ISCA Speech Synthesis Workshop (SSW-7)*, Kyoto, Japan.

[Wood, 1973a] Wood, S. (1973a). Speech tempo. *Working Papers Lund, Phonetics Laboratory Lund University*, IX:144–184.

[Wood, 1973b] Wood, S. (1973b). What happens to vowels and consonants when we speak faster? *Working Papers Lund, Phonetics Laboratory Lund University*, VII:8–29.

[Wrede, 2002] Wrede, B. (2002). *Modelling the Effects of Speech Rate Variation for Automatic Speech Recognition*. PhD thesis, Universität Bielefeld.

[Wu and Sun, 2000] Wu, Y. and Sun, X. (2000). How fast can we really change pitch? Maximum speed of pitch change revisited. In *Proceedings International Conference on Spoken Language Processing (ICSLP)*, Beijing, China.

[Yoshimura et al., 1998] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. In *Proceedings ICSLP 1998*, volume 2, pages 29–32.

[Young et al., 2006] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.-Y., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The Hidden Markov Model Toolkit (HTK) version 3.4. http://htk.eng.cam.ac.uk/. [Online; accessed 16-December-2016].

[Zen et al., 2007] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based speech synthesis system version 2.0. In *Proceedings 6th ISCA Speech Synthesis Workshop (SSW-6)*, volume 2, pages 294–299, Bonn, Germany.

[Zwicker, 1982] Zwicker, E. (1982). *Psychoakustik*. Hochschultext.

[Zwicky, 1972] Zwicky, A. (1972). On casual speech. In *Proceedings Eighth Regional Meeting of the Chicago Linguistic Society*, pages 607–615, Chicago, IL.