Temporal Entrainment in Overlapping Speech

Marcin Włodarczak

Dissertation zur Erlangung des akademischen Grades Doctor philosophiae (Dr. phil.) vorgelegt an der Fakultät für Linguistik und Literaturwissenschaft der Universität Bielefeld am 8. Oktober 2013.

Prüfungskommission:
Prof. Dr. Petra Wagner (Betreuerin und Gutachterin)
Prof. Dr. Mattias Heldner (Gutachter)
Prof. Dr. Dafydd Gibbon
Prof. Dr. David Schlangen

Datum der mündlichen Prüfung: 13 Januar 2014

♾ Gedruckt auf säure-freiem, alterungsbeständigen Papier (nach ISO 9706).

Under the seeming disorder of the old city, wherever the old city is working successfully, is a marvelous order for maintaining the safety of the streets and the freedom of the city. It is a complex order. Its essence is intricacy of sidewalk life, bringing with it a constant succession of eyes. This order is all composed of movement and change, and although it is life, not art, we may fancifully call it the art form of the city and liken it to the dance—not to a simple-minded precision dance with everyone kicking up at the same time, twirling in unison and bowing off en masse, but to an intricate ballet in which the individual dancers and ensembles all have distinctive parts which miraculously reinforce each other and compose an orderly whole. The ballet of the good city sidewealk never repeats itself from place to place, and in any one place is always replete with new improvisations.

But there is nothing simple about that order itself, or the bewildering number of components that go into it. Most of those components are specialised in one way or another. They unite in their joint effect upon the sidewealk, which is not specialized in the least. That is its strength.

Jane Jacobs, *The Death and Life of Great American Cities*

Contents

List of Figures

Kurzfassung

Sprachliches Verhalten im zwischenmenschlichen Dialog ist gelegentlich mit Aktivitäten wie Tanzen oder Sport verglichen worden. Derartige Metaphern sollen verdeutlichen, dass Dialogen zwischen menschlichen Sprechern ein Anpassungsprozess zugrunde liegt, der sich in Koordination im Verhalten der Dialogpartner hinsichtlich zeitlicher Struktur, Form und Inhalt äußert. Die Dialogforschung hat verschiedene linguistische und behaviorale Variablen identifiziert, an denen sich solche Anpassungsprozesse beobachten lassen, darunter Körperhaltung, syntaktische Struktur und die Auswahl lexikalischer Einheiten. Adaption im Dialog wurde darüber hinaus auch für phonetische Eigenschaften des Sprachsignals, wie beispielsweise Grundfrequenz, Intensität, Stimmqualität oder Sprechrate nachgewiesen. Auf der Grundlage dieser Ergebnisse wurden Modelle der Koordination im Dialog entwickelt, die verschiedene Aspekte des Adaptionsprozesses zwischen Sprechern berücksichtigen. Einige dieser Modelle postulieren automatische Prozesse als zugrunde liegende Mechanismen dieser Koordination, während andere von vollständig intentionalen Vorgängen ausgehen.

Ähnliche Vorstellungen sind in der Forschung auch hinsichtlich der zeitlichen Organisation von Redebeiträgen (*Turns*) im Dialog geäußert worden. Es wurde vermutet, dass der Zeitpunkt der Initiierung eines Turns durch einen Sprecher vom Sprechrhythmus des Dialogpartners beeinflusst sein könnte. Die weit verbreitete Vorstellung, dass die Abfolge von Turns zwischen Dialogpartnern zeitlich sehr präzise und nahezu ohne größere Pausen oder Überlappungen ablaufe, kann in solchen Modellen damit erklärt werden, dass sich Dialogteilnehmer gleichsam auf den Sprechrhythmus des Gesprächspartners "einpendeln" und somit einen Zeitrahmen für die eigene Sprachproduktion vorgegeben bekommen. Dieser hypothetische Mechanismus wird als *Entrainment* bezeichnet. Ähnliche Modelle wurden auch für andere Arten von synchronisiertem Verhalten vorgeschlagen und knüpfen an Theorien an, die in rhythmischen Strukturen ein zentrales Organisationsprinzip für Bewegung und zwischenmenschliche Koordination ausmachen. Empirische Untersuchungen haben jedoch bislang wenig Evidenz für die Wirksamkeit derartiger Mechanismen im Dialog erbracht.

In der vorliegenden Studie werden Ergebnisse präsentiert, die diese Forschungslücke füllen. Im Gegensatz zu älteren Arbeiten, in denen zumeist nur Turns mit "nahtlosen" Übergängen untersucht wurden, werden dabei gerade solche Situationen in den Blick genommen, in denen sich die Redebeiträge von Dialogpartnern zeitlich überlappen. Diese Auswahl erlaubt es, die zeitliche Abfolge von Äußerungen verschiedener Sprecher direkt in den Daten zu beobachten, ohne dass, wie in älteren Ansätzen, rhythmische Muster über Pausen hinweg extrapoliert werden müssen. Die Einbeziehung überlappender Turns erschließt darüber hinaus eine wichtige und bislang vernachlässigte Datenquelle, da neuere Untersuchungen ergeben haben, dass dieses Phänomen in natürlichsprachlichen Dialogen weitaus häufiger auftritt als in der älteren Interaktionsforschung angenommen. Zur Quantifizierung der zeitlichen Struktur der Sprecherabfolge bei überlappenden Äußerungen wird einfaches Maß vorgestellt, das die Position der Initiierung eines Turns relativ zu bestimmten Punkten im Turn des vorangegangenen Sprechers angibt.

Zentrales Ergebnis der Untersuchung ist, dass überlappende Turns nicht an zufälligen Zeitpunkten initiiert werden. Vielmehr wird die Initiierung eines Turns durch phonetisch saliente Elemente in der Sprache des Dialogpartners, wie etwa Vokalonsets oder akzentuierte Silben beeinflusst. Das Phänomen tritt in Sprachen mit unterschiedlichen phonologischen Eigenschaften (Deutsch, Englisch, Französisch, Finnisch) auf, so dass von einem universellen, perzeptiv motivierten Effekt auszugehen ist. Stärkere rhythmische Regularität verstärkt den Effekt, was auf Sprechrhythmus als zugrundeliegenden Mechanismus hindeutet. Absolute zeitliche Periodizität von Äußerungen ist jedoch keine Voraussetzung für erfolgreiche Koordination zwischen Dialogpartnern.

Im weiteren Verlauf der Untersuchung werden Zusammenhänge mit anderen linguistischen und interaktionalen Einflüssen aufgezeigt. So treten überlappende Turn-Initiierungen meist am Anfang oder am Ende eines Turns des Dialogpartners auf. Besonders häufig sind Überlappungen außerdem am Ende von überdurchschnittlich langen vokalischen Intervallen, was möglicherweise als Reaktion auf Disfluenzen oder andere Produktionsprobleme des Dialogpartners hindeutet. Wir argumentieren, dass in solchen Fällen von einem reaktiven oder prädiktiven Mechanismus auszugehen ist und keine Evidenz für automatisches Entrainment vorliegt. Dagegen zeigen sich keine Einflüsse von anderen potentiell bedeutsamen Faktoren (pragmatische Funktion der überlappenden Äußerungen, Wortgrenzen, Modalität, Bekanntheitsgrad zwischen Sprechern usw.) auf das Timing von überlappenden Turn-Initiierungen.

Die vorliegende Arbeit ist damit die erste Studie, in der es gelingt, empirische Evidenz für zeitliche Koordination von Redebeiträgen zwischen Sprechern vorzulegen. Unsere Ergebnisse beinhalten darüber hinaus wichtige Implika-

tionen für Modelle der Adaption zwischen Sprechern sowie für Modelle der *Online*-Sprachproduktion und -perzeption im Allgemeinen. Wir argumentieren, dass der beobachtet Effekt ein emergentes rhythmisches Phänomen darstellt und am besten im Rahmen von koordinativen Strukturen in einem aufgabendynamischen Modell beschrieben werden kann, die Beschränkungen für die Handlungen der einzelnen Dialogpartner vorgeben. Damit liefern unsere Ergebnisse greifbare Evidenz für einen Zusammenhang zwischen Wahrnehmung und Produktion: Die Wahrnehmung des Sprechrhythmus eines Dialogpartners beeinflusst die zeitliche Koordination der eigenen Sprachproduktion. Abschließend liefern unsere Ergebnisse auch ein weiteres starkes Argument, überlappende Redebeiträge als wichtiges Element im menschlichen Dialog zu betrachten und dementsprechend in empirische Dialogstudien einzubeziehen.

Die Arbeit ist wie folgt strukturiert: In Kapitel 1 wird zunächst ein Überblick über die Forschungsliteratur zu Adaption und Koordination im Dialog gegeben. Verschiedene Beispiele für Koordination im Dialog werden diskutiert; außerdem werden verschiedene theoretische Ansätze betrachtet, wobei das Hauptaugenmerk auf dynamischer Modellierung liegt. In Kapitel 2 wird das Phänomen der überlappenden Sprache und sein Status in existierenden Dialogmodellen erörtert. Dabei wird ein möglichst repräsentativer Überblick über Forschungsansätze zu überlappender Sprache angestrebt. Kapitel 3 bietet einen Überblick über Rhythmusbasierte Modelle der zeitlichen Koordination von Turns in Dialogen. Insbesondere wird dabei auf die Modelle von Couper-Kuhlen (1993) und Wilson und Wilson (2005) eingegangen. In Kapitel 4 werden die verwendete Methode zur Quantifizierung der zeitlichen Struktur von überlappenden Turn-Initiierungen und die benutzten Sprachkorpora vorgestellt. Die Ergebnisse der Untersuchung werden in Kapitel 5 beschrieben, in Kapitel 6 diskutiert und in Kapitel 7 zu einer abschließenden Betrachtung zusammengeführt.

Acknowledgements

First and foremost, I would like to thank my supervisor, Petra Wagner, who offered me endless assistance and tolerated my frequent forays into many seemingly unrelated scientific areas.

I am greatly indebted to my dissertation committee members: Mattias Heldner, who also acted as a thorough but sympathetic external reviewer, Dafydd Gibbon and David Schlangen, for a challenging and stimulating discussion during the defence.

I would also like to express my deep gratitude to Juraj Šimko, who has been a good friend and an inspiring collaborator. Had it not been for him (and a fair amount of cold *Pils* at the no longer existent *Butzemann*) this work would never have been created.

I thank my Finnish collaborators, Michael O'Dell, Mietta Lennes and Tommi Nieminen, for fruitful collaboration and important methodological suggestions.

I have benefitted greatly from discussions with members of the Bielefeld Phonetics and Phonology Group, Zofia Malisz, Laura de Ruiter, Barbara Samlowski and Andreas Windmann, as well as other colleagues I have had a privilege of collaborating with: Hendrik Buschmeier, Spyros Kousidis, Jens Edlund, Maciej Karpiński and Catha Oertel.

My special thanks go to my friends, especially Krzysztof, Mateusz, Roman, Jagna and Abir, for helping me keep my mind off work and retain (relative) sanity.

Last but not least, I thank my parents, whose constant encouragement allowed me to persevere through many a creative crisis and bring this work to a successful conclusion.

Introduction

Speaking in dialogue has been on some occasions likened to activities such as dancing and playing sports. What such metaphors emphasise is an underlying adaptation process as a result of which dialogue partners' behaviour becomes coordinated in time, form and content. Earlier research has identified a number of linguistic and behavioural dimensions involved in such interspeaker adaptation, including postural sway, syntactic structure and lexical choice. Several reports of adaptation have also been put forward in connection with phonetic features, such as F_0, intensity, voice quality and speaking rate. The findings have given rise to numerous models of interspeaker coordination addressing different aspects of adaptation, from purely automatic to fully intentional.

Related claims have been made about temporal orderliness of *dialogue turns*. Namely, it has been suggested that timing of turn onsets might be influenced by rhythmic properties of interlocutor's speech. On that view, the proposed precision of turn-taking attested by the purported rarity of instances of silence and overlap might be explained by the fact that dialogue participants *entrain to* (or tune into) their interlocutor's speech rhythm, which in turn serves as a temporal frame guiding their own speech. The models fit in well with other known examples of interpersonal synchronisation and with notions of rhythm as an organising principle behind human action and interpersonal coordination. However, so far there has been little evidence in their favour.

In this work we aim to fill this gap by examining timing of *overlapping speech* onsets. We claim that overlapping speech is much more suitable for tracing interspeaker entrainment than "clean" speaker changes, which have been the focus of attention in most earlier studies. Specifically, overlapping speech allows observing temporal relationships between speakers directly in the data without the necessity of extrapolating rhythmic patterns over periods of silence. In addition, overlapping speech has been recently demonstrated to be much more pervasive in human interaction than previously assumed, and its omission in earlier studies might have been a serious oversight. Consequently, we propose a simple measure for quantifying the instant of overlapping speech onset relative to certain landmarks in the previous speaker's turn.

Out results indicate that overlapping speech is not initiated at random locations but is guided by phonetically salient events, such as vowel onsets and prominences associated with pitch accents and duration, in interlocutor's speech. Equivalent entrainment patterns across languages with different phonological properties (English, German, French, Finnish) hint at a universal perceptually-motivated effect. In addition, as more regular duration patterns have been found to facilitate entrainment, the underlying mechanism appears to be rhythm-based. It does not, however, require strict isochrony as a prerequisite for successful coordination and accommodates moderate deviations from perfect periodicity.

In addition, the low-level entrainment to prominences in the flow of speech has been found to co-exist with influences of other levels of interactional organisation. For instance, presence of simultaneous starts and terminal overlaps results in greater concentration of overlap onsets in the vicinity of turn boundaries. Similar clustering has been observed towards the end of lengthened intervocalic intervals, possibly corresponding to overlap being produced in response to interlocutor's disfluencies and other production problems. We claim that in these cases overlap is triggered by reactive or predictive mechanisms and should not be considered evidence of entrainment proper. At the same time, overlap timing remains insensitive to other potentially relevant factors, such as pragmatic function of overlapping utterances, rhythmic entrainment to lexical boundaries or dialogue setup (visual contact, familiarity between speakers, etc.).

The present work provides the long lacking evidence for interspeaker coordination in turn timing. Moreover, it has important implications for models of interspeaker adaptation and, more generally, online speech production and perception. We argue that the observed effect is a rhythm-mediated emergent phenomenon and is best described in terms of task defined coordinative structures, which impose constraints on individual interlocutors' actions. Our findings thus offer a tangible evidence of a link between perception and action, whereby perceiving interlocutor's speech rhythm shapes the minute temporal coordination involved in overlap initiation. Finally, the work provides yet another argument in favour of including overlapping speech as a legitimate part of human dialogue.

The work has the following structure. It opens with a review of interspeaker adaptation in Chapter 1. A number of examples of dialogue coordination are provided alongside an overview of the theoretical accounts of the phenomenon found in literature. Special emphasis is placed on the dynamical modelling paradigm. Chapter 2 deals with the topic of overlapping speech and its status in existing models of turn-taking. An attempt is also made to provide a possibly representative review of research directions related to overlapping speech. Chapter 3 reviews rhythm-based approaches to explaining timing of dialogue

turns. In particular, models by Couper-Kuhlen (1993) and Wilson and Wilson (2005) are discussed in some detail. Chapter 4 introduces the method used to quantify temporal patterns of overlap initiation and describes the corpus resources used. The results are presented in Chapter 5, followed by discussion and conclusions in Chapters 6 and 7.

The thesis is in part based on or re-uses the following previously published material:

- Włodarczak, M., J. Šimko, and P. Wagner (2012). Syllable boundary effect: temporal entrainment in overlapped speech. In *Proceedings of Speech Prosody 2012*, pp. 611–614. (Section 5.1)

- Włodarczak, M., J. Šimko, and P. Wagner (2012). Temporal entrainment in overlapped speech: Cross-linguistic study. In *Proceedings of Interspeech 2012*, Portland, OR. (Section 5.2)

- Włodarczak, M., J. Šimko, P. Wagner, M. O'Dell, M. Lennes, and T. Nieminen (2013). Finnish rhythmic structure and entrainment in overlapped speech. In E. L. Asu and P. Lippus (Eds.), *Nordic Prosody. Proceedings of the XIth Conference*, Frankfurt am Mein, pp. 421–430. Peter Lang. (Section 5.5.4)

- Włodarczak, M., J. Šimko, and P. Wagner (2013). Pitch and duration as a basis for entrainment of overlapped speech onsets. In *Proceedings of Interspeech 2013*, Lyon, France, pp. 535–538. (Sections 5.3 and 6)

- Włodarczak, M. and P. Wagner (2013). Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps. In *Proceedings of Interspeech 2013*, Lyon, France, pp. 1434–1437. (Section 2.1.1)

The analysis and interpretation of the Finnish data in Sections 5.2 and 5.5.4 were done by Michael O'Dell and Juraj Šimko. The remaining analyses are by the present author.

Chapter 1

Interspeaker adaptation in dialogue

The chapter introduces the notion of interspeaker adaption, a phenomenon whereby people who speak (or otherwise interact) with one another, become similar in their behaviour. The effect has been long known to exist but so far lacks a unified theoretical, descriptive or even terminological treatment. As the field is plagued by severe naming confusions, Section 1.1 discusses definitional variation found in literature. Subsequently, a representative body of evidence for interspeaker influence across a wide range of linguistic and interactional dimensions is presented in Section 1.2. A review of theories of interpersonal coordination follows in Section 1.3 with special emphasis on models which attempt to apply principles of dynamical systems theory to human interaction.

1.1 Defining entrainment

Even a cursory review of literature on interspeaker adaptation reveals a multitude of terms used to describe the phenomenon. For instance, Burgoon et al. (2007) list the following: *adaptive responses*, *accommodation*, *interpersonal coordination*, *matching*, *mirroring*, *convergence*, *reciprocity*, *mimicry*, *compensation*, *divergence*, *complementarity*, *synchrony*, *dissynchrony* and *mutual influence*. The inventory is by no means exhaustive and other names, such as *chameleon effect* (Chartrand and Bargh, 1999), exist in literature. To make things even worse, the terminology is rarely used consistently across theories, leading to a situation in which the same phenomenon is known under different names and the same name refers to different phenomena when used by different authors.

In an attempt to organise the field, Burgoon et al. (2007, p. 116–131) proposed the following criteria for classifying adaptive phenomena:

1. Contingency upon interlocutor's behaviour as opposed to dependence on purely individual or interaction-external factors (e.g. mood or social norms).
2. Presence of mutuality of influence.
3. Presence of behavioural change in contrast to maintenance of previous behaviour.
4. Measuring magnitude of change versus measuring direction of change.
5. Presence of temporal and rhythmic dependencies.
6. Intentionality.
7. Requirement for strict equivalence of displayed behaviours.

These led them to offer the following definitions:

Matching: exhibiting behavioural similarity between two interactants, regardless of its cause, intentionality, or partner influence.

Mirroring: behavioural matching in the form of identical, static visual behaviours (e.g. postural asymmetry).

Convergence: the process of interaction adaptation whereby one adopts behaviour that is increasingly similar to that of the partner.

Interactional synchrony: similarity in rhythmic qualities and enmeshing or coordination of the behavioural patterns of both parties.

Reciprocity: adaptation in which one responds, in a similar direction, to partner's behaviours with behaviours of comparable functional value.

Complementarity: exhibiting of dissimilar behaviour, regardless of cause, intent, or partner behaviour.

Divergence: the process of interaction adaptation whereby one adopts behaviours that are increasingly dissimilar from that of the partner.

Compensation: adaptation in which one responds with behaviours of comparable functional value but in the opposite direction.

Matching and complementarity were proposed as the most general concepts, serving as umbrella terms for other types of adaptive behaviour.

Unfortunately, while the above definitions might provide some guidance in the terminological confusion surrounding studies of interpersonal adaptation, they have not been widely adopted by the community, nor has any other consistent naming scheme. For instance, Pickering and Garrod (2004) use *coordination* to refer to Burgoon et al.'s (2007) behavioural matching, which they contrast with *alignment*, consisting in sharing of mental representations. As a result, Kousidis's (2010) conclusion that "the operational definition adopted in each case suits the proposed methodology and theory" is as valid today as ever.

Notably, although entrainment, the central notion of the present work, is conspicuously absent from Burgoon et al.'s (2007) list, it can be subsumed under the rubric of interactional synchrony. It should be noted, however, that entrainment has particularly often fallen victim to definitional inconsistencies. The term has been used to refer to similarity in lexical choice (Brennan and Clark, 1996; Brennan, 1996), phonetic adaptation (Levitan et al., 2011; Levitan and Hirschberg, 2011; Lee et al., 2010) and temporal coordination (Cummins, 2009a). In the present work entrainment is used in the last of these senses. Adaptation and coordination will be used as general terms. A more precise definition of entrainment will be provided in Section 1.3.5.

In addition to Burgoon et al.'s (2007) classification, several convenient distinctions between adaptation phenomena were formulated in the context of Communicative Accommodation Theory (CAT, Giles et al. 1991). While some of them reflect the specific methodological position and focus of CAT itself, others are applicable more generally. First, CAT contrasts *upward* and *downward* adaptation, depending on whether a speaker shifts into a more or less socially prestigious language variety. Second, a distinction is made between between *unimodal* and *multimodal* adaptation, affecting one or several modalities respectively. An important point made by Giles et al. (1991) is that in the multimodal case convergence on some features can be accompanied by divergence on other features. Similarly, when convergence on a specific feature is impossible (e.g. specific dialectal variety not shared by interlocutors), speakers might compensate for it by adopting an analogous feature from their own repertoire (e.g. by switching to their own local dialect). Furthermore, *symmetrical* processes, in which both parties move towards or away from each other, are contrasted with *asymmetrical* ones, in which only one party modifies his or her behaviour. Depending on the amount of (mis)matching adaptation can be *full* or *partial*, with the additional possibility of *hyperconvergence*, which, however, is normally perceived negatively by interlocutors. Two final distinctions within CAT are between *conscious* and *unconscious*, and between *subjective* and *objective* processes. Importantly, percep-

tual bias might mediate adaptation behaviour both in speakers and in listeners. On the perception side, objective adaptation, as measured by some physical parameter, may deviate from speakers' subjective beliefs about their partners' and their own adaptation practices. In extreme cases objective convergence can thus be subjectively perceived as divergence. On the production side, speakers have been observed to adapt to what they *believe* their interlocutors' behaviour is. Similarly, convergence is sometimes directed towards some prototypical goal expected of the speaker by the interlocutor, implying the role of social stereotypes in shaping perception of dialogue partners.

1.2 Evidence for interspeaker adaptation

Interspeaker adaptation in dialogue has been reported for a great number of linguistic and paralinguistic features from fine phonetic detail to lexical and syntactic forms to temporal patterning of whole dialogue turns. Although by no means fully consistent with each other, the findings provide strong empirical evidence for the existence of interpersonal adaptation. This section attempts to present a possibly wide spectrum of methodologies, results and interpretations.

1.2.1 Lexical and syntactic adaptation

On the lexical level, Garrod and Anderson (1987) demonstrated that speakers in a maze task collaboratively develop shared description schemes. According to the authors, such schemes allow users to arrive at common situation models, and are established through *input/output co-ordination*, which consists in adopting the most recent description used by the interlocutor. By contrast, Brennan and Clark (1996) proposed that speakers end up using similar terms not merely due to recency effects but as a result of jointly and provisionally established dyad-specific *conceptual pacts*, agreements about object conceptualisations, which are reinforced by their frequent use.

The relationship between lexical adaptation and interaction quality was investigated by Nenkova et al. (2008), who found that adaptation on words with highest frequencies in a corpus is indicative of (perceived) interaction naturalness and task success. It is also positively correlated with overlap frequency and negatively correlated with the number of interruptions, which the authors interpret as evidence of engaged and coordinated turn-taking behaviour. Niederhoffer and Pennebaker (2002) reported similarly high correlation between frequencies of various word classes (e.g. words greater than 6 letters, prepositions, words referring to positive emotions) used by dialogue partners, but found no relation between the degree of matching and participants' reports

on interaction quality. The latter finding led the authors to hypothesise that adaptation is related to conversational involvement rather than interpersonal rapport. At the same time, unlike Ward and Litman (2007), who evaluated repetition rate of lexical items correcting for topic and lexicon (i.e. lack of synonyms) constraints and using a baseline of randomised turns, Niederhoffer and Pennebaker (2002) employed no random baseline. Neither did they attempt to eliminate effects of topic, which could potentially alter the results for at least some word classes, such as past tense verbs.

Syntactic structures have been also found to be sensitive to interpersonal influence. Branigan et al. (2000) reported evidence of syntactic priming on the realisation of ditransitive verbs in a variant of a picture description experiment. The degree of adaptation was higher when the same lexical form was used. Gries (2005) obtained compatible results for priming of dative alternation and particle placement in a corpus-based study. He additionally identified a link between the strength of syntactic preference for a particular construction of individual verbs and their susceptibility to priming. Finally, Reitter et al. (2006) investigated the amount of syntactic priming in spontaneous and task-oriented dialogues across all possible constructions by counting repetitions of syntactic patterns at different distances from the primes. He concluded that more priming occurs in task-oriented than in spontaneous dialogues, as well as within than across speakers. Additionally, low-frequency constructions were found more likely to be primed, and the probability of priming to decrease logarithmically with the distance from the prime. More recently, the validity of findings on syntactic priming was questioned by Howes et al. (2010), who compared the degree of syntactic similarity in dative constructions found in spontaneous interactions against a baseline of dialogues constructed from randomly selected turns and sentences. They found little evidence supporting occurrence of syntactical similarity, and concluded that the regularities observed might be largely due to reusing of the same lexical forms.

1.2.2 Phonetic adaptation

Adaptation on phonetic parameters has been explored using a plethora of methods (both experimental and corpus-based) and metrics for capturing interdependencies between speakers. For instance, Levitan and Hirschberg (2011) demonstrated adaptation between dialogue partners along four phonetic dimensions (intensity, F_0, voice quality, and speaking rate) both globally, over whole conversations, and locally, on a turn-by-turn basis, in *proximity* (similar feature values), *convergence* (gradual decrease of differences) and *synchrony* (mirroring of *relative* change in one speaker by the other). On the whole, speakers were

more similar to their dialogue partners than to randomly chosen speakers with whom they did not interact. Lee et al. (2010) investigated adaptation in F_0 and intensity slopes calculated over 100 ms time windows between adjoining turns using a combination of signal processing and information theoretic methods. Their results indicate higher adaptation rate for positive-outcome interactions (the effect was stronger for pitch-derived features). The same features were successfully used to discriminate between interactions with highly positive and highly negative attitudinal content using Markov modelling. Convergence of pitch, intensity and speech rate was also confirmed by Kousidis et al. (2009) using bivariate time series analysis. Yet another method was used by Buder et al. (2010), who found prosodic coordination in mother-infant interactions using cross-recurrence analysis and coupled-oscillators models. Periods of convergence and divergence were found to coincide with phases of mutual engagement and disengagement of the parties respectively. Similar results were obtained for multiparty conversations by De Looze et al. (2011) using correlation-based measures of prosodic similarity calculated over a series of overlapping windows. Finally, Ward and Litman (2007) employed the technique proposed by Reitter et al. (2006) for syntactic priming (see Section 1.2.1) by tracing the likelihood of matching F_0 and intensity excursions in neighbouring turns. Significant relationships were recorded for both features.

In contrast to the signal-based studies discussed above, Pardo (2006) assessed phonetic similarity in map task dialogues with ABX testing. According to the author, a perceptual method is superior for determining phonetic adaptation as it allows a global evaluation of similarity and does not concentrate on adaptation along individual features, which are likely to vary between speakers and dyads (Pardo, 2012). In her study each speaker's map task landmarks produced early or late in the conversation were compared against their partners' productions and their own readings of target phrases one to two weeks before and directly after completing the task. Target phrases were judged as more similar to partners' production than to own pre- or post-task readings with stronger effect for items produced late in the task, evidencing convergence between dialogue partners. Phrases produced during the task were also perceived as more similar to post-task items than to pre-task items, indicating a degree of persistence of accommodation phenomena. Due to presence of a statistically significant effect of speaker's gender and role on accommodation patterns, the author challenged accounts of alignment based on priming and the direct perception-production link, and postulated that the degree of coordination must be instead modulated by factors external to the perception-production system. Accordingly, the author claimed that phonetic alignment might be more appropriately modelled as entrainment between coupled dynamical systems (see Section 1.3.5).

Phonetic adaptation has also been observed for specific dialogue phenomena. In particular, Levitan et al. (2011) showed adaptation in phonetic realisations of backchannel-preceding cues (Gravano and Hirschberg, 2011). Statistically significant correlations between mean pitch and intensity values in neighbouring cues from different speakers testified to a local nature of the coordinative mechanisms. The degree of interspeaker influence was also demonstrated to be positively correlated with task success and better coordination in backchannel timing. In a similar vein, backchannels were shown to be closer in pitch to dialogue partner's directly preceding speech than were other utterance types (Heldner et al., 2010).

Notably, phonetic adaptation can be mediated by cultural and social factors as well as expectations of the dialogue partner. Welkowitz et al. (1972) found evidence of amplitude adaptation only between those of the randomly assigned dialogue partners who had been told they were paired with a similar interlocutor. No effect was observed for participants who were aware of the random pairing. In addition, Lewandowski (2012) demonstrated that the degree to which non-native speakers of English converged on amplitude envelopes to their native English interlocutors is positively related to their pronunciation talent.

1.2.3 Temporal adaptation

Much of the work on temporal adaptation in dialogue goes back to Jaffe and Feldstein's (1970) pioneering work on speech and silence sequencing in dialogue. They found durations of silences and vocalisations to be stable within as well as between conversations with the same dialogue partner and on the same topic, regardless of the time elapsed between individual interactions. Additionally, durations of silences but not of vocalisations were influenced by different dialogue partners, and the amount of overlapping talk was constant within individual dialogues but varied between conversations. At the same time, stable patterns characteristic of individual speakers (speaker styles) were also observed.

While later studies generally replicated Jaffe and Feldstein's (1970) findings regarding silences, conflicting results were reported for durations of vocalisations. For example, Street (1984) demonstrated convergence in male-male dyads but divergence in male-female dyads. Cappella and Planalp (1981) also found some indications of similarity in vocalisation patterns but in their data the moment-by-moment durational matching was in some dyads countered by compensatory tendencies. The authors also concluded that the influences in question are generally weak, especially when compared with within-speaker

consistency. In addition, an association between durations of interlocutor's vocalisations was observed for specific dialogue act categories, such as questions and answers in an interview setting (Matarazzo et al., 1963).

High correlation of *average* between-speaker pause durations were also reported by ten Bosch et al. (2005). However, as demonstrated by Kousidis and Dorran (2009) interspeaker convergence is difficult to track on a moment-by-moment basis and might be overridden by dialogue state or utterance type. A related claim was made by Edlund et al. (2009), who suggested that convergence and synchrony in dialogue should be modelled in a dynamic fashion rather than by means of crude measures comparing averaged values across two parts of an interaction. Edlund et al. (2009) calculated correlations between interlocutors' silence durations in a moving window and found some evidence of speaker coordination. Crucially, similar to Kousidis and Dorran (2009), the results of the frame-by-frame analysis were not necessarily in line with the results obtained with global measures. Their findings are thus also in line with those studies of phonetic adaptation (see Section 1.2.2) which identified *periods* of interspeaker similarity but no overall coordination patterns (e.g. De Looze et al., 2011).

Convergence on response latency and speech rate in fact-finding interviews was examined by Street (1984) using time series regression. As expected, the degree of similarity influenced participants' judgement about each other's social attractiveness. Importantly, evidence for convergence was much stronger when all data was pooled than in individual dyads, suggesting a subtle nature of the effect, which can be easily obscured by factors such as topic changes. Similarly, Finlayson et al. (2011) found that map task participants' speech rates become increasing similar across a conversation, and are influenced by their interlocutors' speech tempo in the previous turn. Ward and Nakagawa (2004) investigated speech tempo adaptation in directory assistance dialogue with a view to improving naturalness of dialogue systems. They found that telephone numbers were dictated at slower speeds to slow speaking users, who took longer to react to the operator's greeting, and at higher speeds to fast speaking and fast reacting users. Ward and Mamidipally (2008) attempted to reproduce these results in a more complex billing support domain. However, only slight correlations between tempi of neighbouring utterances were found, and factors related to subtask and utterance type were also found to jointly determine speaking rates.

Similar to other cases of adaptation, temporal matching has been claimed to be associated with positive speaker rating and good interaction quality. Warner et al. (1987) found that moderate coordination between dialogue partners' vocal activity elicits higher ratings in terms of affect, involvement, relaxation, etc. than cases of more extreme matching. They also found stronger synchronisation

between interlocutors' vocalisation patterns than between individual's vocal activity and their own heart rate, suggesting that behaviour might be conditioned predominantly by social constraints and only to a lesser extent by physiological variables. Natale (1975a,b) found that the degree of interspeaker convergence on switching pause durations and speech intensity levels is greater in speakers with high social desirability scores. Importantly, in his study speakers tended to converge on pause duration during the second interaction only, which indicates that this type of adaptation indeed develops gradually over time, as suggested by Jaffe and Feldstein (1970).

1.2.4 Multimodal adaptation

Coordination of body movement was first investigated by Condon and Ogston (1971). They reported tight temporal matching between speech and body parts movement both within a single speaker and across dialogue partners. Subsequently, Kendon (1970) concluded that the auditory channel alone is sufficient for gestural synchronisation. These results, however, were based on examination of selected interaction episodes and lacked a rigorous quantification. More recently, Shockley et al. (2003) and Shockley et al. (2007) used cross-recurrence analysis to demonstrate that coordination of postural sway between partners involved in the same collaborative task can indeed be mediated by speech and does not require that interlocutors see each other. Relatedly, Altmann (2010) measured movement synchrony in interactions between children using windowed cross-lagged regression.

Building on a body of work suggesting existence of a direct priming-based *perception–action* link in social behaviour and social judgements formation, Chartrand and Bargh (1999) proposed that the same unconscious and fully automatic mechanism underlie behavioural mimicry. On that view, simply watching an action increases the likelihood of similar behaviour in the observer. They demonstrated that participants tended to copy the behaviour of confederates they interacted with, such as smiling, rubbing one's cheek and shaking one's foot. The effect appeared not to be goal-dependent (e.g. in terms of seeking social acceptance) or conscious but nonetheless resulted in participants' increased likeability and interaction smoothness.

Richardson et al. (2007) found evidence of gaze coordination between people discussing images presented on a shared screen. The degree of matching was dependent on participants' common ground: it was higher when both heard the same background information before the task. Additionally, Richardson et al. (2008) discovered that *beliefs* about common ground also influence gaze patterns. Specifically, greater gaze coordination was observed when participants looking

at separate screens thought they did not share the visual scene, in which case they employed additional verbal grounding strategies.

A potentially fertile but much understudied domain is coordination of respiratory activity. The field was pioneered by McFarland (2001), who reported synchronous patterns (both in-phase and anti-phase) of interlocutors' breathing cycles near turn transitions. Unfortunately, to the best of this author's knowledge, it has been the only systematic study of respiratory matching in dialogue to date.

1.2.5 Adaptation in human-computer interaction

Interestingly, adaptation on all the levels discussed above has also been elicited in experiments involving human-computer interaction. Stoyanchev and Stent (2009) and Branigan et al. (2003) presented evidence of syntactic priming in text-based human-computer interaction. Since in Branigan et al.'s (2003) experiment adaptation occurred regardless of whether participants were told they would be interacting with a computer or a human partner (in reality interactions were scripted in all cases), the authors concluded that adaptation is independent of speakers' beliefs about their interlocutors' mental states. Consequently, the findings were interpreted as evidence in favour of automatic models of alignment (see Section 1.3). However, as rightly pointed out by Brennan (1996), who reported that users also tend to copy system's lexical choices, matching computer's productions might be a conscious strategy aimed at minimising comprehension errors.

Zoltan-Ford (1991) found that users' inputs tend to match the length and vocabulary of system prompts. However, the effect was weaker when the system used longer, conversational prompts. No difference was observed between text- and speech-based interfaces with regard to adaptation phenomena. Similarly, Bell et al. (2000) found limited support for an effect of interface type used for system output generation (point-and-click or speech) on users' choice of modality in system-addressed referring constructions.

Bell et al. (2003) investigated speech rate adaptation in interactions with a virtual agent. While there was an overall increase in speech rate as dialogue progressed, the effect was stronger among participants interacting with fast rather than slowly speaking agent. At the same time, agent's speaking rate had no influence on users' silent pause durations. Finally, amplitude convergence was reproduced in a human-computer setting by Coulston et al. (2002), who found that 7-10-year-old children adapted their intensity levels to those of a virtual agent they interacted with.

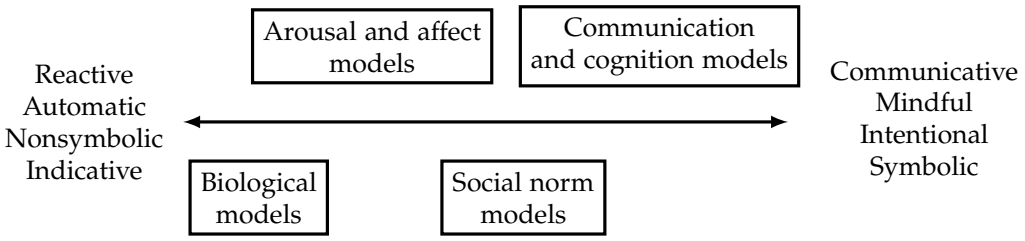

Figure 1.1: Classes of interpersonal adaptation theories arranged on a reactive-communicative continuum (adapted from Burgoon et al. 2007).

1.3 Models of interspeaker adaptation

Given the amount of empirical evidence, only a fraction of which was presented above, the existence of interspeaker adaptation is difficult to deny. It appears to be pervasive in communication between humans but is also present in interactions with dialogue systems and avatars. By contrast, interpretations of mechanisms underlying the observed adaptation patterns depend greatly on authors' methodological and theoretical stance, which has given rise to many, partly conflicting, accounts.

A particularly convenient classification of interpersonal adaption models has been proposed by Burgoon et al. (2007), who in their comprehensive review of literature described a continuum going from purely biologically-based theories focusing on automatic and mostly innate patterns of behaviour to high-level models, which aim at accounting for communicative functions of adaptive patterns. According to this criterion, theories can be grouped into four hierarchical categories comprising *biological models*, *arousal and affect models*, *social norm models* and *communication and cognition models*. The continuum is portrayed schematically in Figure 1.3. Each category is summarised in turn below; subsequently, an alternative view is introduced in the form of embodied theories of entrainment drawing on ideas from dynamical systems theory. While the latter have a lot in common with Burgoon et al.'s (2007) biological class, especially the interactional synchrony models, they offer a powerful apparatus applicable to much wider domains of human interaction.

1.3.1 Biological models

Biologically-motivated theories of adaptation focus on automatic and unintentional adaptive patterns. In particular, much effort has gone into studying synchronous behaviour, conceptualised in terms of congruence of interlocutors' *behavioural rhythms*, *simultaneous behaviour* or *perceptual synchrony* (Bernieri

et al., 1988), especially in connection with coordination of speech and body movements (Condon and Ogston, 1971). Such models posit existence of certain *interactional rhythms* underlying the observable synchronisation of behaviour. Importantly, interactional synchrony does not necessarily entail identical behavioural *forms* but merely their relation in time: "[i]nteractional synchrony is defined by [...] isomorphism of pattern of change between the speaker-hearer" (Condon and Ogston, 1971, p. 159). Although synchrony can be modified by a range of cultural or individual factors, it appears to be rooted deeply in human biological make-up and is observed already in very early infancy. The reported rapidity of synchronisation processes also attests to its automatic nature. As such, synchrony is thought to serve basic survival needs such as safety or bonding. However, it has also been posited to help manage interaction flow, signal affective states and facilitate speech processing.

Related biologically-rooted phenomena are those of *motor mimicry* and *mirroring*, that is "the tendency to imitate others' nonverbal expressions, particularly expressions such as laughter, pleasure, embarrassment, pain, discomfort, and physical exertion" (Burgoon et al., 2007, p. 25). Specifically, motor mimicry is said to occur when one reacts to a *stimulus* directed towards someone else, for example by expressing fear when observing someone in a dangerous situation. It has been hypothesised to be both an emphatic response and to indicate cognitive adoption of another person's perspective. By contrast, mirroring refers to copying or imitating interlocutor's *behaviour*. Similar to synchronous behaviour, both mimicry and mirroring are said to facilitate bonding and are sometimes speculated to signal affiliative messages.

Not listed by Burgoon et al. (2007) but very prominent in recent years is Pickering and Garrod's automatic alignment model (AAM, Pickering and Garrod, 2004). Their approach is somewhat different from other models listed in this category because of its strong representational standpoint. AAM defines alignment as "sharing representations at some level" (Pickering and Garrod, 2004, p. 172) and contrasts it with purely behavioural *coordination* in joint activity. In fact, according to AAM dialogue partners are able to coordinate their actions *because* their mental representations become aligned. In other words, it is assumed that successful communication requires sharing of at least some representations between interlocutors, even in unresolved arguments or partial misunderstandings. Nevertheless, AAM aims at drawing a fully mechanistic account of language processing in dialogue, free from resorting to inferences about interlocutors' goals or intentions.

The basic claim of AAM is that alignment of interlocutors' linguistic representations and situation models is mediated through an automatic and resource-free priming process. On this view, perceiving a certain expression activates its asso-

ciated mental representation, thus increasing the probability of subsequently producing a similar expression and leading to formation of a tight coupling between production and perception. In addition, alignment is assumed to spread between levels of linguistic representations, such that alignment on one level (e.g. lexical) leads to alignment on other levels (e.g. syntactic). Finally, Pickering and Garrod (2004, p. 176) propose that priming processes form *channels of alignment*, bidirectional direct and automatic links between corresponding levels of representations in the speaker and in the listener, allowing processing-free coupling between them. The precise nature of the mechanism, however, is not explained.

Automatic alignment is claimed to greatly reduce processing costs in three fundamental ways. First, it eliminates the need for a fully specified common ground allowing dialogue partners to rely instead on shared representations. Second, it leads to emergence of fixed expressions constructed and adopted by interlocutors as interaction develops, and third, it provides a mechanism for self-monitoring and self-correction on each level of representation. Importantly, although fully automatic, alignment in AAM nevertheless requires that dialogue participants *attend* to their interlocutors. It is also proposed that alignment can be inhibited or promoted depending on congruence of interlocutors' goals.

1.3.2 Arousal and affect approaches

Stemming from the study of proxemics, arousal and affect approaches extend biological models by incorporating factors related to interlocutors' psychological needs. Special emphasis is placed on arousal reactions triggered by violating intimacy expectations of dialogue partners, and the resulting responses aimed at reinstating equilibrium. In early formulations, such as Affiliative Conflict Theory (ACT, Argyle and Dean 1965), all deviations from a state of equilibrium were assumed to lead to compensatory behaviour. As this leaves no space for reciprocity and matching, later theories have attempted to account for the ebb and flow between the two forces. For example, in a modified formulation of ACT, Argyle and Cook (1976) introduced social and functional factors acting as triggers of non-compensatory reactions. Reciprocity and compensation have received a more extensive treatment in Arousal Labelling Theory (Patterson, 1976), which posits that arousal reactions are labelled (evaluated) positively or negatively and trigger corresponding behavioural responses. Similarly, Markus-Kaplan and Kaplan's Bidimensional Model (Markus-Kaplan and Kaplan, 1984) accounts for both tendencies by introducing two dimensions of *individuation – deindividuation* (corresponding broadly to the strength of one's ego) and *attachment – detachment* (corresponding to the strength of the relationship).

Participants' reactions to intimacy changes depends on their placement in this two-dimensional space, with interactions between individuated, attached types producing reciprocity, and interactions between deindividualised, detached and individuated, attached types leading to compensation. Finally, Capella and Green's Discrepancy-Arousal Theory (Cappella and Green, 1982, 1984) proposed a direct and automatic link between arousal and affect. Its direction depends on the magnitude of the deviation of dialogue partner's behaviour from one's expectations and acceptance regions: small deviations within the acceptance region produce positive affect and approach reactions while large deviations falling outside of that region produce negative affect and avoidance reactions. As in the Bidimensional Model, reciprocal and compensatory patterns are produced by the combination of individuals' affective reactions: approach-approach and approach-avoidance respectively.

1.3.3 Social norms models

Social norms theories comprise a class of models which emphasise the role of socially-motivated mechanisms behind interpersonal adaptation. Possibly the most prominent of these is Communication Accommodation Theory (CAT, Giles et al. 1991), which started with the aim of explaining accent convergence in interviews but has since been expanded to account for other adapation phenomena. Interestingly, because of its of broad scope, it is claimed to be capable of accounting for interspeaker adaptation from a micro moment-by-moment adjustments up to the level of global language changes.

CAT defines convergence as "a strategy whereby individuals adapt to each other's communicative behaviours in terms of a wide range of linguistic-prosodic-nonverbal features" (Giles et al., 1991, p. 7). It is a means of associating with an interlocutor and is considered to fulfil the need for social integration, social approval and identification with a dialogue partner (even though the need itself need not be conscious). Convergence works in accordance with the principle of *similarity attraction* (Byrne, 1971), whereby perceived similarity increases likeability and encourages positive evaluation of the speaker. Indeed, the body of work reviewed by Giles et al. (1991) provides evidence that convergence results in interlocutor's increased attractiveness, competence and communicative effectiveness. Furthermore, perception of convergence can be mediated by social norms as well as intentions and the amount of adaptive effort attributed to an interlocutor, with more effort resulting in more favourable evaluation. Importantly, according to CAT social benefits of convergence need to be offset against its costs related to personal or social identity loss, the effort involved and possible unfavourable interpretations, such as being perceived as less intelligent.

By contrast, divergence, defined as "a way in which speakers accentuate speech and nonverbal differences between themselves and others" (Giles et al., 1991, p. 7), is a tactic of distancing oneself from an interlocutor and seeking social identity in terms of identification with a different social group than that of a dialogue partner. For example, divergence can be used to stress speaker's ethnic membership when it is challenged by an interlocutor. Similarly, switching to a standard language variety when talking with a speaker of a local dialect is perceived as emphasising one's privileged social status. However, such divergence strategies are normally evaluated negatively by interlocutors.

Notably, CAT introduces an *inter-group* perspective on what are normally considered to be purely *interpersonal* phenomena, thus bringing in a set of group-related variables shaping the extent of convergence/divergence, such as the perception of own and partner's group prestige. Obviously, apart from socially-driven accommodation, adaptation on purely personal grounds is also possible but, as Giles et al. (1991) point out, the two types of factors usually cannot be easily disentangled.

In addition to the functions outlined above, both convergence and divergence might assist cognitive organisation of events, thus catering for interlocutors' communicative needs. For instance, adapting to a non-native speaker's simplified syntax might be a strategy of compensating for their limited language proficiency. Similarly, divergence can be used in drawing interlocutor's attention to one's own lacking language skills by emphasising a foreign accent, or in an attempt to slow down a fast speaking interlocutor by adopting a markedly slow speaking rate. This is an entirely different perspective on interspeaker accommodation, which reinterprets it in terms of sensitivity to particular speaker's conversational needs rather than seeing it as a mere "approximation strategy" (Giles et al., 1991, p. 41) to the perceived patterns of a dialogue partner.

Also included by Burgoon et al. (2007) in the social norm class are theories grounded in Gouldner's norm of reciprocity (Gouldner, 1960) and social exchange theory (Roloff, 1987). These approaches consider reciprocity in the form of exchange of resources (both material, such as money, and nominal, such as information, affection and social status) as the basis of achieving and maintaining stability both at the level of individual relationships and in a society at large. Similar in vein but stressing harmful aspects of accommodation are reports of reciprocation in couple interactions, especially in couples facing problems in their relationships and seeking counselling.

1.3.4 Communication and cognitive models

The final category of adaptation models listed by Burgoon et al. (2007), the communication and cognitive approaches, offers a functional perspective on adaptation and interprets such phenomena in the context of participants' communicative goals. Adaptation patterns are viewed as conveying certain interactional meanings, whose recognition and interpretation by interlocutors determines their behavioural responses. As a rule, these models try to explain both automatic and non-automatic processes. For example, in his Sequential-Functional Model, Patterson (1983) presented an account of how pre-interactional factors, such as personality traits, past experiences, automatic behavioural patterns and affective assessments combine to form a framework for interpreting functions of interlocutor's involvement behaviours. Each action is evaluated in the context of its (possibly multiple) inferred communicative functions and compared against behaviours expected by the addressee. Similar to the models discussed previously, behaviours which are in line with their inferred function and interlocutors' expectations are evaluated positively and lead to stable exchanges characterised by reciprocity. By contrast, large disparities between the observed and the expected behaviours lead to negatively-evaluated high arousal levels and to compensation. Similar assumptions underlie the Interaction Adaptation Theory (Burgoon et al., 2007), in which each participant's behavioural needs, socially-motivated expectations and personal goals jointly determine, possibly with different weights, individual *interactional positions*, which are then compared against interlocutor's behaviour, and adaptations occur towards the more positively evaluated of the two.

By contrast, Cognitive-Valence Theory (Andersen, 1992) posits that only moderate levels of arousal trigger conscious cognitive valencing. In those cases interlocutor's behaviour receive positive valence if all aspects of that behaviour (cultural, relational, interpersonal, situational, etc.) are evaluated positively. As before, high arousal levels are directly linked to negative valence, and low arousal levels to positive valence. Lastly, in a reformulation of the Motor Mimicry Theory, Bavelas et al. (1986) reinterpret motor mimicry as a primarily *communicative act* of care, sympathy and rapport towards the interlocutor.

1.3.5 Dynamical theories of entrainment

Burgoon et al.'s (2007) classification is a coherent and well-structured attempt at explicating assumptions behind theoretical approaches to interpersonal adaptation. It proposes a hierarchy of models which account for increasingly complex mechanisms behind coordination patterns. We close this chapter by discussing a radically different perspective on coordination adopting instead the apparatus

of dynamical systems theory and coordination dynamics (Kelso, 1995). It abandons many of the traditional notions about cognition, language and interaction, and offers a unified treatment of traditionally unrelated phenomena, such as motor coordination, formation of joint action, empathy, etc.

The method has been successfully used in physical sciences and is in many respects much more suitable for explaining coordination phenomena observed in speech and in dialogue. It is particularly well suited for explaining emergence of stable *temporal patterns* between interacting systems. What follows is a review of the fundamental concepts and their application to interpersonal interaction.

Dialogue and coordinative structures

Coordinative structures were originally employed to explain motor coordination in blacksmiths by Bernstein (1967). He speculated that the number of muscles involved in performing even simple tasks is far too great to be controlled centrally by the brain. Instead, he argued, coordinative structures (or synergies) emerging in the context of a particular task (e.g. hitting an anvil with a hammer) group degrees of freedom into task-specific functional units. These task-specific groupings impose local constraints on movement, effectively reducing system complexity compared to the summed complexity of the component parts. Bernstein (1967) also noted their extraordinary and instantaneous compensatory abilities evident in the perfect coordination of hammer movement despite variance in hand trajectory.

The same mechanism has been subsequently instrumental to modelling motor coordination in other domains. Notably, coordinative structures provided a satisfactory account of movements of articulators in speech production (Kelso et al., 1980). Here too the inherent complexity of the task and the great number of degrees of freedom make explanations in terms of direct control exercised by the brain implausible. Similarly strong propensity for compensation has also been observed: physically constraining one articulator triggered an instantaneous readjustment of other articulators in reaching a specified target.

Coordinative structures can thus be conceived of as leading to spontaneous and local "emergence, maintenance, and dissolution of special-purpose kicking machines, scratching machines, speaking machines, throwing machines, etc." (Cummins, 2011a, p. 4) in which many complex parts unite in the service of a shared behavioural goal. The resulting systems exhibit strong tendencies for periodicity behaving analogously to non-linear oscillatory mass-spring-like systems and can be conveniently described by coupled oscillators models. Given their cyclic tendencies, emergence of constraints on movement within coordinative structures leads to mutual *entrainment* of the component systems with their periods or relative phases of oscillation becoming coupled.

More recently, several authors have proposed that similar coordination mechanisms might govern interpersonal coordination. Specifically, it has been demonstrated that coordinative structures are by no means restricted to the domain of an individual. Indeed, since coordinative structures are defined functionally, on a task basis, they are able to bridge the chasm between seemingly irreconcilable domains and transcend "the somewhat arbitrary boundaries separating brains, bodies and environments" (Cummins, 2011a, p. 5). From that point of view, speakers can be modelled as autonomous dynamical systems coupled by the joint task (Cummins, 2009a). At the same time, as Marsh et al. (2009) rightly point out, crossing the interpersonal divide requires that the same dynamical principles operating within a single individual be found in interpersonal interactions, and that the links be strictly informational.

This possibility was investigated in the experiment conducted by Schmidt et al. (1990). The researches modified the original experiment by Haken et al. (1985), in which subjects were asked to move both their index fingers cyclically with constant frequency. Haken et al. (1985) found very strong constraints on the relative phase of oscillations, with only two stable values: one in which both fingers moved in-phase or were anti-phased. Schmidt et al. (1990) used a similar task but distributed it among two people by asking them to swing one of their legs while seated. Similar constraints on frequency and phase were identified, indicating that visual contact is sufficient for the emergence of synchronised action. Since then a number of examples of spontaneous interpersonal entrainment have been reported, for example between hand-swung pendula (Schmidt et al., 2007) or between people sitting in rocking chairs (Richardson et al., 2007). Coordinative structures have also been posited as the mechanism underlying gaze and body sway coordination between dialogue partners (Richardson et al., 2009; Shockley et al., 2009, 2007; Richardson et al., 2007). Importantly, in the experiment by Shockley et al. (2007) body sway entrainment was found to be mediated by speech. Even more relevant to speech research were experiments by Cummins (2009b), who showed that completely untrained subjects are extremely skilled at synchronising their speech when reading an unfamiliar text. The precision they achieve, measured in terms of asynchronies between the respective speech signals, is commonly within the range of 40-60 ms with further training having little effect on their performance. Additionally, the mismatch between speakers does not accumulate across the task, hinting at compensatory mechanisms similar to those observed in physical systems. Emergence of interpersonal coordinative structures with similar dynamical properties has also been observed in cooperative social action with solo and joint action acting as attractors (Marsh et al., 2009). Indeed, it has been suggested that synchrony might lie at the heart of social connection and that "being pulled into the orbit of

another's body may be a rather fundamental, body-based way of instantiating a socioemotional connection with another" (Marsh et al., 2009, p. 334).

Furthermore, the reduced dimenisonality of the coordinated system has been proposed as a powerful and parsimonious account of the ease and naturalness of interaction and the ubiquity of adaptation between interlocutors. Moore (2012) argued from an information theoretical point of view that synchronisation offers gains due to increased mutual information between components and lowering information rate of the resulting system. It has been also suggested that complementarity at higher levels of organisation in dialogue might serve a similar purpose of reducing the overall system complexity. According to Fusaroli et al. (2013), phenomena such as speaker change and adjacency pairs should not be discussed in terms of normative static "scripts" but in terms of emergent interactional routines setting limits on the interactional dynamics. In the same vein, Rączaszek-Leonardi (2009) and Rączaszek-Leonardi and Kelso (2008) argued that symbolic representations, traditionally regarded as fully static, should be conceived of primarily in terms of emergent dynamic constraints on communication.

One of the consequences of generalising the notion of coordinative structure into the interpersonal domain is that compensatory patterns similar to those present within a single physical system should be discernible:

> if the cross-person coordinative structure consists of a certain relation among body segments and cognitive states, then constraints on the (action) effectors of one person should affect the movements and/or cognition of the other member of the pair as readily as cognition can affect one's own effectors. (Shockley et al., 2009, p. 315)

This is indeed what is observed in the synchronous speech task: a reading couple appears to have a phenomenal capacity for compensating for mismatches in relative timing even though no consistent separation into leader and follower roles is present. By contrast, a serious reading mistake on the part of one speaker results in *both* speakers stopping immediately. Likewise, compensatory tendencies have also been reported for speakers with speech impairments (Dressler et al., 2009).

Entrainment and the role of rhythm

Whatever the channel of mutual influence (physical/informational), there must exist some mechanism resulting in entrainment between autonomous parts. Motivated by the oscillatory tendencies of coordinative structures mentioned in the previous section, Cummins (2009a) proposed that the link lies in the underlying rhythmicity offering affordances for the entrainment of movement.

On this view, rhythm leads to emergence of stable temporal patterns in movement of otherwise independent systems. Crucially, rhythmicity as proposed by Cummins (2009a) is a property of neither a stimulus nor of the reaction to that stimulus. Instead, it is precisely what allows coordination of the two. In other words, rhythm is a *relation* "between the capacity for movement of a specific person and the structure (temporal or otherwise) of a specific signal" (Cummins, 2009a, 18).

Notably, even though the channels of influence might be strictly informational, the underlying movements are always physical (whether they be movements of clock pendula, fingers, articulators or torsos). Consequently, a theory of rhythm as an affordance for movement is necessarily an embodied one, in that it requires accounting for properties of physical objects interacting with each other in a physical environment. Needless to say, human interactions, both verbal and nonverbal, fall into precisely that category.

Nevertheless, Cummins (2009a) points out that rapid changes and aperiodicity found in speech (and, by extension, in human interactions in general) pose certain problems for coupled oscillator models (Barbosa, 2002; O'Dell and Nieminen, 1999) commonly used to describe behaviour of entrained systems. At the same time, coordination need not necessarily result in perfect periodicity. In fact, Buder (1991) demonstrated how a simple fully deterministic non-linear model of speaker involvement is capable of accounting for a wide range of dynamical patterns, including asymptotic stability, periodic oscillation and chaotic behaviour. Furthermore, contrary to most accounts of rhythm in speech, neither does the speech signal itself need to exhibit obvious, easily observable regularity:

> If the performance of two synchronous speakers is accurately described as coupling among dynamical systems, there must be a basis for that coupling. That is, there must be some means for one system to influence the other. Presumably this information is to be found primarily in the speech signal, but a cautionary note is in order here. The principle of affordance expresses a fit between the information available in perception to an organism, and the action possibilities available to that organism. There is no part of this description that requires that the physical description of that information be simple in the sense of readily decomposable by analysis into "basic" physical variables. (Cummins, 2009a, 25)

Indeed, as shown by Inden et al. (2012), oscillators can be successfully used to model rhythmic structure of spontaneous speech.

The hypothesis of rhythm acting as an affordance for the entrainment of movement was subsequently (Cummins, 2011b) refined employing the account

of sensorimotor coordination as a *skilled action* originating in the work of Dewey (1896) and his criticism of the sequential processing model implicit in the concept of the reflex arc. Following Dewey, Cummins defines skilled action as "the imposition of constraints on the co-variation of movement and sensory flux such that the boundary conditions that define the skill are met" (Cummins, 2011b, p. 1). In other words, performing any skilled action involves coordination of perception and action within some predefined context without a fixed, linear succession of the two. Synchronisation with an external stimulus is thus tantamount to integrating that stimulus into one's own production such that within a context of a particular skill they are perceived as one and "collectively function as the sensory arc of the sensorimotor coordination" (Cummins, 2011b, p. 5). For instance, in synchronous speech this amounts to overlaying of speech of the other speaker onto one's own speech and treating it as one speech signal. This view of synchronisation with an external stimulus allows for loosening of the requirement for strict periodicity as a prerequisite for entrainment. Additionally, it emphasises the coordination of movement and perception in establishing coordinative structures.

Cummins is by no means alone in proposing rhythmicity as the organising principle behind coordinated action. Other authors have also suggested that rhythm might facilitate motor coordination and social behaviour both within and between organisms, often appealing to notions of endogenous or internal, physiologically-motivated rhythms. For instance, Chapple (1970) conjectured that endogenous rhythms determine and modulate much of individual and interpersonal behaviour. Warner (1979) found regular alternation between periods of high and low speaking activity with cycles of 3 and 6 minutes long, possibly related to regularity in breathing patterns due to speech-induced periods of hyper- and hypoventilation. Entrainment of respiratory activity itself was subsequently reported by McFarland (2001). At a much larger time scale, Hayes and Cobb (1979) suggested that readiness to engage in interaction is determined by an underlying endogenous biorhythm with the period of roughly 90 minutes, similar to rhythms found in dreaming activity during sleep. In addition, it has been suggested that rhythmic entrainment to interlocutor's speech might serve as a basis for alignment of higher-level linguistic representations (Wagner et al., in press).

1.4 Conclusions

The present chapter offered an overview of examples and modelling paradigms of interpersonal coordination. As suggested by Figure 1.3 these range from purely reactive mechanisms of mimicry to coordination described in terms

of intentional communicative acts. As a consequence, theories closer to the latter end of the continuum incorporate increasingly many assumptions about interlocutors' psychological needs and expectations (often conceptualised as, possibly automatic, affective judgements of their dialogue partner's behaviour), social pressures shaping the interaction flow or inferences about intentions and meanings behind the observed behaviour. Associated with the growing theoretical baggage is weak support of these models in empirical results. While ample evidence exists for the presence of coordination at the biological level, this is less true of the remaining models. Clearly, the reason does not so much lie in their limited theoretical appeal but rather in the intrinsic difficulty of testing their highly abstract claims.

Furthermore, the models inherit many of the traditional views on cognition, action and perception. Indeed, even when the posited underlying coordination mechanisms are fully automatic priming-like processes, they rely strongly on mental representations, whether social, behavioural or linguistic (Bargh and Chartrand, 1999; Pickering and Garrod, 2004). Accordingly, perception and production become coupled in each interlocutor not because of some functional interdependence between them but because they draw from the same representational inventory. By the same token, at the interpersonal level interlocutors are hypothesised to exhibit similar behaviour due to their situation models becoming increasingly alike. The real locus of alignment is, therefore, on representational rather than behavioural level, a point made very strongly by Pickering and Garrod (2004). As a consequence, dialogue partners remain autonomous monads with production and perception processes which are independent of each other both within and across speakers and which simply use the same representational resources. In addition, as pointed out by Fusaroli et al. (2013), a fully automatic, indiscriminate priming mechanism should lead to attempts at perfect imitation of the dialogue partner, which is commonly perceived negatively by dialogue partners (Giles et al., 1991) and associated with poorer task performance (Fusaroli et al., 2012).

A radically different perspective on interspeaker adaptation is offered by dynamical models. Perhaps most importantly, it is

> agnostic about the locus of agency, and focus[es] instead on domains within which lawfulness may be found in the spatio-temporal change over time of observables. These domains may be transient in nature, they may cut across the boundaries of nervous system, bodies, and environments, and they may be defined over multiple individuals as well as within a single organism. (Cummins, 2011b, 2)

Neither does this account depend on the notion of mental representations, relying instead on physically-motivated models of coordination. As Shockley et al. (2009) point out such an approach to modelling coordination

> does not accord a special status to cognitive representations. They are constraints upon coordination just as any other and can be analyzed in the same language used for physical systems. For example, the type of spontaneous alignment described by Garrod and Pickering is not special to cognitive linguistic systems, and can be observed in many places in physics and biology where there is a tendency for units to coordinate (i.e., pulling coupled units into coordinated modes).

Finally, dynamical models of entrainment provide a plausible account of effects such as synchronous speech. These results obviously pose a serious problem for traditional models of speech perception and production since they are at odds with delays introduced by centrally managed error recovery (Cummins, 2011a). Neither can priming or a perception-behaviour link (Bargh and Chartrand, 1999) provide a satisfactory explanation for what is a case of synchronisation and not of mirroring or copying interlocutor's behaviour, especially in light of aperiodicity of speech and no strict division into leader/follower roles.

Obviously, the perspective on interaction sketched out above is only possible when human beings are not modelled as autonomous information-processing units building internal representations of the outside world. Fusaroli et al. (2013) argues that adopting a dynamic perspective necessitates rejecting two long-standing assumptions about dialogue: first, that the ultimate goal of conversation is mutual understanding achieved through alignment of mental representations and second, that participants are fully autonomous agents. Instead, participating in dialogue should be regarded as a joint activity in which each party can be studied only as a part of the emerging organisation.

Differences between the dynamical approach and models such as the Automatic Alignment Model should now become clear. On the coordinative structures account, interacting individuals are not independent entities who arrive at similar hierarchical mental representations through mutual priming. Instead, the observed coordination is a result of spontaneous organisation of the interacting parties, such that they form a unit with new (and simpler) properties defined by speakers' intentions, task-related and physical constraints, etc. (Shockley et al., 2009). Fusaroli et al. (2013) formulate the character of coordination succinctly as a specification of

> how local task requirements come to guide and constrain alignment and, even more importantly, distribute *complementary* (rather than

identical) actions among interlocutors making them temporarily coupled, selective aligned and fulfilling different roles in the interaction

The nature of this organisation is not essentially different from that found in all physical systems and reflects their tendency to remain in or oscillate around states characterised by low energy and entropy (Moore, 2012). For this reason, models built around the notion of the coordinative structure are necessarily embodied as they are concerned with physical agents acting in physical environments. As Marsh et al. (2009) point out, from the dynamical perspective the predominant functions of the brain are not those of "directing behavior but instead biasing the system toward selecting certain environment-evoked behaviours in one's behavioral repertoire over others." In other words, cognition is seen primarily as imposing constraints on organism's behaviour (Shockley et al., 2009).

Finally, since coordinative structures are defined on a functional basis, their emergence and characteristics are dependent on the task at hand. This stands in sharp contrast to priming-based approaches in which alignment is an automatic consequence of perceiving a stimulus or otherwise requires a decision to suppress it, which in turn incurs certain processing costs. The same property of coordinative structures might also explain lack of consistent patterns across studies of interpersonal coordination. Different experimental set-ups or participants' traits might simply lead to formation of coordinative structures with somewhat different properties and dynamics.

Chapter 2

Overlapping speech in dialogue

The present chapter focuses on overlapping speech, the other central theme of the present work. While the topic has been present in dialogue literature since its early days, overlaps have been often treated as a marginal or aberrant phenomenon, and it has been only recently that they have been considered worthy of attention and a more systematic investigation by a wider community. As a consequence, it seems fit that the overview offered below should attempt to combine some of the historical views on overlap with the state-of-the-art findings in the field. This dual approach is particularly called for in the light of the lasting impact of the early misconceptions, which haunt the field to this day.

As overlapping speech lies at the heart of the problem of establishing speaking order, the chapter opens with an overview of major turn-taking models in Section 2.1 with special emphasis on treatment of overlaps in each framework. Formal, rule-based models are subsequently contrasted with approaches which seek to explain overlap solely in terms of contextual factors such as stylistic or contextual variation (Section 2.2). The chapter closes with a necessarily fragmentary review of the by now large body of findings related to the multiple aspects of overlapping speech in human-human and human-machine interaction.

2.1 Overlaps in turn taking systems

Wilson et al. (1984) classified turn-taking models into three categories based on their methodological and theoretical assumptions: *stochastic*, *signalling* and *sequential*. They are discussed briefly below, alongside the more recent *turn-bidding* model. In particular, the status granted to overlapping speech in each class of models is examined in some detail.

2.1.1 Stochastic models

Stochastic models go back to the method of *interaction chronography*, first used by Norwine and Murphy (1938) to describe temporal patterns of speech and silence in telephony conversations. The fundamental notion in their approach was that of a *talkspurt*, i.e. "speech by one party, including his pauses, which is preceded and followed, with or without intervening pauses, by speech from the other party perceptible to the one producing the talkspurt" (Norwine and Murphy, 1938, p. 282). Vocalisation segments were inferred automatically from the audio as five second signal samples exceeding a pre-defined intensity threshold. Signal samples falling below this threshold were classified as silence. A similar method was also used by Chapple (1939) to study speakers' personality traits.

The technique was subsequently refined by Jaffe and Feldstein (1970) to include unilateral *vocalisations*, *speaker switches*, *pauses* (silences bounded by vocalisation of the same speaker), *switching pauses* (silences bounded by vocalisations of different speakers) and *simultaneous speech*. This classification allowed them to model temporal patterns of speaking in dialogue stochastically as a first-order Markov model defined in terms of four *dyadic states*: unilateral vocalisation by each of the speaker, simultaneous vocalisations, and simultaneous silence. A further development was proposed by Cappella and Planalp (1981) by including a history of past transition probabilities in each speaker's model.

While objectivity of chronography-based segmentation is advocated as one of its major advantages, it is not completely free from theoretical assumptions. Specifically, although the models themselves are fully deterministic, their underlying elements (vocalisations and silences) are derived from signal intensity within a specified time window. Even though the window length is claimed to be dictated by perceptual threshold on silence and speech detection, there is little systematic evidence supporting the exact value chosen. For example, Jaffe and Feldstein (1970, p. 18) set the threshold to 300 ms, corresponding "to the natural common sense perception of sound burst and pause in speech." Recently, Włodarczak and Wagner (2013) have demonstrated that silence boundary thresholds do indeed have a substantial effect on durations and frequencies of various within- and between-speaker intervals.

Crucially, the stochastic approach does not treat stretches of overlapping speech as special, neither theoretically nor descriptively. The reason for this is that *speaker change* as such is conceived of in purely probabilistic terms, devoid of any semantic or pragmatic meaning. According to Wilson et al. (1984) "[t]he phenomenon of speaker change is, of course, provided for in the stochastic model, but its status is simply that of a transition between abstractly defined states that occur with probabilities estimated directly from the data for a given conversation" (Wilson et al., 1984, p. 162). To gain further insights into turn-

taking behaviour vocalisation patterns need to and have been successfully combined with information about location of other 'hidden events' which locally modify transition probabilities. Jaffe and Feldstein (1970) themselves identified an increased probability of speaker change following "linguistically permissible phrase endings" (Jaffe and Feldstein, 1970, 49). Other authors found high likelihood of interruption onsets after disfluencies, discourse markers, backchannels and filled pauses (Shriberg et al., 2001). Stochastic modelling has also been used to investigate the dynamics of overlap initiation and resolution in multi-party conversations (Laskowski et al., 2012, see Section 2.3.7).

2.1.2 Signalling models

In signalling models turn-taking mechanisms in conversation are assumed to be governed by a set of cultural rules coordinating participants' behaviour. The rules are "mediated through signals composed of clear-cut behavioral cues, considered to be perceived as discrete" (Duncan, 1972, p. 283). In other words, signals indicate participants' turn taking states and are exchanged to achieve a 'smooth flow' of conversation. While similar claims have been made by various authors (e.g. Yngve 1970), signalling models are most commonly associated with the work of Duncan (1972) and Duncan and Fiske (1977).

Unlike in the stochastic approaches, the segmentation method employed by Duncan and Fiske (1977) was purposefully subjective[1] with units corresponding to 'phonemic clauses' bounded by events such as phrase final pitch movement, grammatical completion, audible exhalation or relaxation of the foot. Somewhat surprisingly, the authors admitted that "the reasons for deciding to define the units of analysis on the basis of the occurrence of the listed actions cannot be fully recalled" (Duncan and Fiske, 1977, p. 168).

Duncan and Fiske's (1977) turn-taking system essentially consists of turn-taking signals, rules describing possible actions at a specific point in an interaction and actual moves (e.g. turn take or release). The authors compiled a non-exhaustive list of multimodal cues associated with, among others, turn yielding, turn holding and backchanneling based on their high co-occurrence with the respective actions. Additionally, it was found that combinations of many features are more effective in achieving their desired goal than combinations of few features, a finding reproduced recently by Gravano and Hirschberg (2011).

In a system like that of Duncan and Fiske (1977) overlapping turns can arise in a number of ways. For instance, dialogue participants might grab a turn

[1] "The procedure was designed to reflect subjective impressions of the points at which it seemed reasonable to draw unit boundaries" (Duncan and Fiske, 1977, p. 174)

which has not been explicitly relinquished, or they might continue to speak in spite of displaying a turn yielding cue. Similarly, presence of both turn-yielding and turn-holding signals results in an increased probability of simultaneous speech. However, unless overlapping speech is of a very specific type (e.g. backchannels, overlaps with filled pauses, audible inhalations or *sociocentric sequences* such as "you know"), it indicates the turn taking mechanisms has "broken down" or has been "discarded", and is (to be) avoided:

> Just as it desirable to avoid bumping into people on the street, it is desirable to avoid in conversation an inordinate amount of simultaneous talking. Beyond considerations of etiquette, it is difficult to maintain comprehensibility when participants in a conversation are talking at the same time. (Duncan, 1972, p.283).

Importantly, overlap resolution (i.e. deciding who continues after overlap termination) is not within the scope of the turn taking mechanism itself. No turn-resolution system was proposed by the authors, though.

2.1.3 Sequential models

Sequential models originate in the field of Conversation Analysis (CA) and have found their most representative formulation in the classic model of Sacks et al. (1974). The authors start by making a number of observations about properties of turn-taking in spontaneous conversation pertaining to turn allocation and turn timing (Sacks et al., 1974, p. 700), most famously:

> "Overwhelmingly, one party talks at a time"
>
> "Occurrences of more than one speaker at a time are common, but brief"
>
> "Transitions (from one turn to a next) with no gap and no overlap are common. Together with transitions characterized by slight gap or slight overlap, they make up the vast majority of transitions"

They also note variability in turn size, order, semantic content, etc., as well as existence of discontinuous speech and repair mechanisms.

Subsequently, Sacks et al. (1974) go on to propose a model, which is *locally managed* (allocating speakership on a turn-by-turn basis), *interactionally managed* and characterised by *recipient design*, i.e. sensitive and oriented towards a particular interlocutor. The model comprises two components specifying permissible turn-constructional units and turn allocation mechanisms. According to Sacks

et al. (1974) turns consist of units which have the property of projectability, that is which "allow a projection of the unit-type under way, and what, roughly, it will take for an instance of that unit-type to be completed" (p. 702). Projectability is defined mainly in syntactic terms but other levels of linguistic organisation (such as prosody) might also contribute to end-of-turn prediction. The first completion point of such a unit becomes a *transition relevance place* (TRP) at which turn ownership is managed according to three rules: selection of the next speaker by the previous speaker, self-selection of the next speaker or continuation by the previous speaker. The rules apply sequentially in the order in which they are listed; in other words, other participants may self-select unless the previous speaker has designated the next speaker, and the previous speaker may continue unless some other participant has self-selected. If the previous speaker has continued with his or her turn, the rules are re-applied at the next TRP. Otherwise, if none of the options has been exercised, the rules for self-selection by another party and continuation by the previous speaker are recycled, eventually resulting in an extended gap or a *lapse*. Importantly, the fact that speaker change is only permissible at (projectable) TRPs coupled with the ordering of the rules imposing time constraints on their application has the effect of minimising occurrences of gaps and overlaps. This results in another significant property of the model, namely that silence is considered to be a product of the turn-taking mechanism rather than a turn-yielding cue (Wilson et al., 1984, p.169). Thus, turn ends are *predicted* rather than *reacted* to.

Nevertheless, in real-life conversations overlaps (as well as silences) do occur. Sacks et al.'s (1974) systems provides two contingencies for their existence. First, in multi-party conversation more than one party might self-select. Most commonly, the first person to self-select gains the turn and other parties simply terminate their turns in progress, which is in itself the simplest repair mechanism. On some occasions, however, second-starters might be permitted to continue if their turn is (and can be recognised as being) of a specific kind, for example addressing comprehension problems. Second, a possibility for brief overlaps is left open because of limits on TRP projectability due to turn-final decrease of speech tempo or addition of optional elements, such as address terms. Indeed, Sacks et al. (1974) claim that words such as "well", "but", "so" are commonly found at turn onsets to avoid the proper utterance content from being overlapped. In a similar vein, Jefferson (1973) asserts that address terms in turn-final position are designed to be overlapped without sacrificing utterance content, should the next speaker misjudge the timing of a TRP. However, as these overlaps are located around TRPs, their resolution is similar to cases of simultaneous starts and consists in the previous speaker simply reaching the end of his or her turn and ceasing to speak.

Similar to Duncan and Fiske (1977), Sacks et al. (1974) did not put forward any special overlap resolution mechanisms. These were incorporated into the model explicitly by Schegloff (2000) as a set of practices for accommodating the sequential organisation of conversation with interlocutors' "outside-turn-taking interests", which might result in occurrences of overlapping speech. Therefore, overlap resolution mechanisms are of "second-order" in the sense that they are used to deal with problems within the turn taking organisation itself. Accordingly, Schegloff maintains strongly that even though overlapping talk does occur in dialogues, "it is co-constructed by reference to one-party-at-a-time as it targeted design feature, rather than to any value, or no value at all" (Schegloff, 2000, p. 3). Indeed, it is for this reason that overlaps can be recognised as "problematic", with a few exceptions such as terminal overlaps located in the vicinity of TRPs, overlaps involving backchannels and "conditional access to the turn" (e.g. assisting the interlocutor in a word search) as well as "choral" talk (e.g. collective greetings).

The remaining overlap types (most of which, according to Schegloff (2000), involve two people) are dealt with by the *overlap resolution device* composed of three elements specifying the resources used in overlap resolution as well as locations and the interactional organisation of their employment. The resources listed by Schegloff include various "hitches and perturbations," such as increased volume and pitch, modification of speech rate, repetitions and termination of one's speech. They are employed by participants in an interactive and online fashion on a syllable-by-syllable basis (as opposed to the turn-by-turn organisation of the "first-order" system), meaning that each overlapping party uses overlap resolution resources in response to the resources used by their interlocutor in the previous syllable. At any stage, a speaker can either drop out of overlap or move to *competitive production*, which might ultimately become an extended *floor fight*. Such longer stretches of competition, however, are claimed to be rare, with one of the speakers usually withdrawing after the first overlapped syllable or after the first syllable produced by their interlocutor in the competitive mode. Effectively, very few overlaps last longer than three syllables and those which do are motivated by speakers' urgency to speak immediately, for example with a view to addressing understanding problems. Additionally, Schegloff (2000) identified six *phases of overlap* (*pre-onset*, *post-onset*, *post-post-onset*, *pre-resolution*, *post-resolution*, *post-post-resolution*), which modify the ways in which overlap resolution resources are used (e.g. increasing or decreasing of speech tempo). Many of Schegloff's (2000) hypotheses have been confirmed by quantitative analyses (Laskowski et al., 2012; Gravano and Hirschberg, 2012).

The resulting system shares the three fundamental characteristic of the turn-taking system proposed by Sacks et al. (1974). It is organised locally on a syllable-by-syllable-basis, party-administered, interactionally-managed and recipient-designed.

2.1.4 Turn bidding model

The turn bidding model of turn-taking was proposed by Selfridge and Heeman (2009, 2010) in the context of dialogue systems as an alternative to turn-yielding moves as the only mechanism of speaker selection. This approach draws on Schegloff's account of motivations for floor fights in terms of urgency to speak. Indeed, the authors explicitly claim that their approach was inspired by practices employed in turn-conflict resolution and that the same strategies operate in all other types of turn transfer. In the turn bidding model conversational parties do not passively react to or predict turns ends. Instead, they weigh the urgency (or importance) of their own message relative to the turn-state cues of their dialogue partner and start speaking prior to their interlocutor's turn completion should their contribution be urgent enough. As turn-conflict resolution might be costly in interactional terms, the possibility of overlap occurrence is claimed to enter the decision of whether to bid for a turn.

The corollary of these assumptions is that "people, wishing to speak, only limit their contributions due to insufficient conversational importance" (Selfridge and Heeman, 2009), and, conversely, the more important an utterance, the stronger the turn-bidding cues used. For the same reason, utterance importance should be negatively correlated with pause duration, as more urgent utterances are likely to follow turn offsets more quickly. On that view, silence duration becomes a turn-bidding cue in its own right. In the authors' view, the proposed system of turn-taking does justice to the truly mixed-initiative character of spontaneous dialogue, in which participants are free to start speaking at any stage if what they are going to say is of sufficient importance.

The model was tested in an artificial collaborative dialogue task between the system and two kinds of simulated users: expert, always allowing short gaps between turns and novice, allowing longer gaps. The system, bidding for the turn after every utterance, was able to adjust to the two types of users, and to match turn-bidding strength with utterance importance.

The turn bidding model offers an alternative perspective on overlapping speech. In conventional models if overlaps were given any special theoretical status at all, they were considered to violate the established turn-taking rules, and special "second-order" repair or resolution strategies were put forward to bring interaction back on its orderly and organised track. In other words,

different rule sets were proposed to account for turn transfer accompanied by silence and overlap. Selfridge and Heeman (2009, 2010) invert the picture by claiming that "smooth" speaker changes are governed by essentially the same mechanism as those operating in overlap resolution. Admittedly, turn bidding was implicitly incorporated into Sacks et al.'s (1974) model in the form of granting first-starters access to the floor. However, as the rules in Sacks et al.'s (1974) system are applied sequentially, the right of the previous speaker to select the next always precedes self-selection. Conversely, the turn bidding model distributes floor control equally among all dialogue parties and posits the mechanisms found in overlap resolution as the fundamental mechanism of speaker change. Overlaps thus becomes a legitimate citizen of a turn-taking system and "provide a window into the inner working of turn-assignment" (Selfridge and Heeman, 2009, p. 2)

2.2 Ecological approaches to overlap

The signalling and sequential approaches characterised above attempt to describe turn management in terms of a set of rules governing smooth transfer of the speaker role. The rules are defined in abstract and general terms, insensitive to conversational context or other pragmatic aspects of communication, and overlapping speech is generally characterised as a breach of conversational laws. Indeed, it has been suggested that systems such as that of Sacks et al. (1974) should be regarded as fundamentally *prescriptive* (O'Connell et al., 1990). While the stochastic models make no such claims, this comes at the price of disregarding virtually all aspects of conversation not related to timing of talkspurts. The bidding approach is somewhat more situationally oriented but still fails to account for factors other than utterance importance. By contrast, in this section we discuss approaches which, for lack of a better term, we refer to as ecological to emphasise their focus on environmental and contextual factors conditioning turn exchange.

In their critical assessment of the existing turn-taking models, O'Connell et al. (1990, p. 346) claim that

> [t]he ultimate criterion for the success of a conversation is not the 'smooth interchange of speaker turns' or any other prescriptive ideal, but the fulfilment of the purposes entertained by the two or more interlocutors.

It is not clear whether authors of any of the models discussed in the previous section would consider efficient turn management "the ultimate criterion" for conversational success but Sacks et al. (1974) do indeed consider turn-taking to

be "the basic form of organization for conversation" (p. 700). In other words, organisation of turn exchanges is what makes certain events of talk by two or more parties recognisable as conversations.[2] The organisation in question is characterised without reference to contextual factors, such as identities of speakers, speech content and other situated aspects. Although Sacks et al. (1974) do maintain that, paradoxically, the system is also sensitive to conversational context, this seems to be mainly related to the dimensions the system leaves unspecified, such as turn-size, turn-order, conversation length, etc.[3] By contrast, the turn allocation resources and the associated avoidance of gaps and overlaps are proposed as universal. Duncan and Fiske (1977, p. 235) are more careful on this point claiming that there should be "no reason to assume a priori that appropriate observance of the turn-system rules is the most important aspect of a conversation" and suggesting that turn-taking can plausibly be assumed to co-exist with other levels of organisation.

O'Connell et al.'s (1990), however, go much further and maintain that no prescriptively ideal system of turn allocation is either a sufficient or a necessary condition for success in human interaction. They criticise the metaphor of conversation as a game in which only one party has ball possession at any time, and hold that conversational time is owned jointly by all interlocutors, whose active participation cannot be limited to production of speech but needs to include other modalities as well. In doing so, they emphasise the contractual character of conversation established jointly by all parties:

> Most of the contractual aspects are implicit, determined by who the interlocutors are, why they are engaging in conversation, how much time they have at their disposal, where the conversation takes place, and many other *contextual* factors. Contracts are *not* devices, but purposeful, sometimes quite habitual, means adopted to attain some conversation goal. If the contract dictates short turns, long pauses, interruptions, overlap, shouting, or any other verbal or nonverbal means, they are *eo ipso* appropriate. (O'Connell et al., 1990, p. 366)

[2] Insofar as this statement is correct, it opens the way for classifying any interaction which does not adhere to Sacks et al.'s (1974) model as belonging to some other kind of a speech-exchange system.

[3] Cf.: "Recipient design is a major basis for that variability of actual conversations glossed by the notion 'context sensitive'. In referring to the particularizing operation of recipient design on turn-size and turn-order, we are noting that parties have ways of individualizing some 'this conversation'; their collaboration in turn-allocation and turn-construction achieves a particular ordering of particular-sized turn and turn-transition characteristics of the particular conversation at a particular point in it" (Sacks et al., 1974, p. 727). It is not clear what other "turn-transition characteristics" are meant here though.

Consequently, conversation cannot be considered a homogeneous phenomenon and temporal characteristics of turn allocation will necessarily vary depending on speakers' goals, cultural norms, politeness and many other factors. Indeed, the authors suggest that obeying the rules proposed by Sacks et al. (1974) might in some contexts be perceived as highly inappropriate.

Along similar lines, Tannen (1994) contrasts two interaction styles: *high involvement* style, characterised by short between-speaker pauses and overlaps, and *high considerateness* style with longer pauses between turns and a preference for avoiding overlap. Her analysis found that in the former style overlapping speech is by no means disruptive and might in fact indicate high engagement and interest in interlocutor's speech. Tannen (1994) cites a number of cultural and situational contexts in which starting one's speech before the end of dialogue partner's turn is perfectly acceptable or even expected. In her view, overlapping speech is perceived as disruptive mainly as a result of clashes between speaking styles and differing expectations for turn onset timing. Overall, she finds Sacks et al.'s (1974) system to reflect "ideology more than than practice" (Tannen, 1994, p. 62).

Points made by authors like O'Connell et al. (1990) and Tannen (1994) are certainly valuable as reminders that overlapping speech has a rightful place in human dialogues and its disruptiveness is conditioned by a wide range of situational variables with complex interactions between them. Where these approaches fall short is systematic, quantitative evidence supporting what are mostly anecdotal claims. O'Connell et al. (1990) readily accept that and call for more research into contextual factors mediating turn-taking mechanisms but given the multitude of factors to be included, ecologically sound investigation of overlap is extremely difficult at best. As a result, most studies of overlapping speech, whose review is offered in the following section, largely ignore the situational factors alluded to above. Nevertheless, the issues of context-dependence and situatedness will recur regularly in the remainder of this chapter. In particular, it will be seen that a misguided way of studying the influence of arbitrarily selected high-level variables (e.g. gender, familiarity) and disregarding all other possibles effects and interactions is likely to lead to entirely misleading conclusions.

2.3 Selected directions of overlap research

2.3.1 Overlaps and interruptions

So far overlapping speech has been taken to mean simply speech produced simultaneously by at least two speakers with some instances being problematic

and disruptive. No terminological distinction though has been made between overlapping speech and *interruptions*, and no attempt to list the characteristics of overlapping speech which make it intrusive. As it turns out, the solution to neither of the issues is straightforward. In addition, the two are highly interrelated as the answer to the question of what makes certain cases of overlapping speech disruptive naturally depends heavily on which instances are considered disruptive in the first place.

However, while intuitively appealing, the distinction between benign overlaps and intrusive interruptions is nowhere near clear, even if certain kinds of simultaneous speech (e.g. backchannels) are excluded from the analysis from the start. The difficulty is clearly reflected in the many differing definitions of interruption. For example, Zimmerman and West (1975) define interruptions in terms of their remoteness from a possible TRP, but consider both interruptions and overlaps problematic, and call them "violations" and "errors" respectively. At the same time, O'Connell et al. (1990) quote Oreström (1983) as referring to both types as acceptable to the current speaker. Yang (1996, p. 1) lists three conditions for classifying simultaneous speech as an interruption: "intention of the main speaker to continue, entrance of the other person into the conversation and disruption or stopping of the main speaker, at least temporarily." However, he then goes on to confuse the reader by dividing interruptions into *competitive* and *collaborative*, depending on their disruptiveness. Yang (1996) himself attributes the distinction to French and Local (1983), who define interruptions in terms of competition for *immediate* turn possession. Similar definitions were adopted by Wells and Macfarlane (1998) and Kurtić et al. (2013). Notably, as these authors work within the CA framework, they demand that simultaneous speech be classified as interruptive only if interlocutors orient to it as such. A related classification of affiliative and disaffiliative overlaps was proposed by Makri-Tsilipakou (1994) by combining the CA methodology with the notion of face-saving and face-threatening strategies (Brown and Levinson, 1987). Somewhat problematically, most of the utterance types she analyses (e.g. topic shifts, backchannels) were assigned to both classes reflecting the fact that they can be used with opposing functions. Makri-Tsilipakou (1994, p. 409–410) claimed further that both aspects can co-exist within a single utterance, and that the same utterance can be more or less affiliative in different contexts. Murata (1994) differentiates between overlaps and interruptions based on the intention of the incoming speaker to deliberately start speaking before the interlocutor is finished. She subdivides interruptions into cooperative, assisting the interrupted speaker, and intrusive, related to topic-changing, floor-taking or expressing disagreement. She notes that cultural factors might determine positive or negative evaluation of intrusive interruptions. Finally, Yang and Heeman (2010)

follow a different route and discuss interruptions in terms of initiative conflicts, operationalised as forward-looking dialogue acts (Allen and Core, 1997) moving conversation in diverging directions. This definition is consistent with earlier claims that interruptions are likely to result in topic shifts (Murata, 1994, p. 386)

Clearly, the confusion here is twofold. The first one is purely terminological and consists in some authors using "interruption" as a generic term for simultaneous speech (cf. Meltzer et al. 1971, p. 392: "by interruptions we mean two persons vocalizing at once"). The second is more complex as it pertains to the problem of which types of overlapping speech should be considered interactionally problematic (cf. Meltzer et al. 1971, p. 392: "It would be a mistake, however, to infer that each interruption event is a miniature battle for ascendance"). It is not entirely certain to what extent the definitions based on TRP proximity, manifest turn competition, initiative conflicts, face strategies and intentionality are congruent (but see below). Moreover, the difference between collaborative and disruptive overlaps is further blurred by the fact that, even though locally disruptive, interruptions might nevertheless be essential for maintaining mutual agreement within longer interactions (Yang, 1996). Another important point was made by Jefferson (1984), who differentiated between intrusiveness and accountability. She pointed out that even though certain instances of overlapping speech are indeed interruptive, many of them are nevertheless initiated at very specific locations within the ongoing turn (see Session 2.3.7) and utilise legitimate resources provided by the turn-taking system. Jefferson (1984) concludes:

> These variously generated onsets can be seen to be at least systematic, if not perfectly 'proper', reasonable, legitimate, rightful, etcetera. And with these orderliness a mass of overlapping talk is lifted from the realm of non-systematic, perhaps unaccountable, perhaps only interactionally-motivated/accountable 'interruption'.

Finally, interruption is by no means restricted to cases of overlapping speech as speakers may intrude into their interlocutor's turn-internal silences. Indeed, Çetin and Shriberg (2006) presented some evidence that speakers tend to perform turn-grabbing utterances during pauses in their dialogue partner's speech.

The discussion is perhaps best summarised by Schegloff (2002), who regards interruption as a *vernacular term of complaint* rather than a technical descriptive category. Moreover, he maintains that interruption is a *category bound action*, i.e. its recognition is conditional on whether or not agents belong to certain categories. As a consequence, men may be found to interrupt women more frequently because certain actions are recognised as interruptions only when performed by men. Crucially, such actions and the categories they are bound to reinforce each other. Very similar claims were made by Tannen (1994), who

argued that interruption is an interpretive (*emic*) category and one casting moral judgement.

The definitional problems go even further and affect interpretation of interruption outcomes. Specifically, once interruptions are defined in terms of turn competition and fighting for floor rights, it becomes natural to introduce the notion of success conceived of in terms of who drops out of overlap first ("loser" or "yielder") and who survives it ("winner"). However, as pointed out by Schegloff (2000), such militaristic definitions of success are far from exhaustive. He argued that on many occasions speakers can be seen to achieve their goals even though they terminate their turns first. For instance, speakers might only want to persist until their utterance is complete or until they have managed to express the gist of their message. They might also yield their turn to steer the following interaction in a specific direction. This complexity of defining success has been recognised by some. For instance, Beattie (1982) and Beňuš et al. (2011) combine the success and completeness criteria to distinguish between overlap, interruption and butting-in.

Notably, despite the considerable differences between definitions of interruption employed by different researchers, when the category is used in corpus-based studies to investigate prosodic and linguistic marking of interruptive speech, the overall results are strikingly similar. The similarity indicates that, perhaps not surprisingly, the various definitional criteria mentioned above are non-orthogonal and capture much of the same variability observed in human interaction. They are also linked to the same resources employed in turn competition, some of which are discussed in the following section.

2.3.2 Resources for turn-competition

Schegloff's (2000) account of overlap resolution lists a number of "hitches and perturbations" employed in overlap resolution. In addition, there is considerable literature which investigates the means of turn-competition in greater detail, with most studies concentrating on prosody, fluency and timing of overlapping speech.

The first systematic exploration of resources used to signal turn competition was that of French and Local (1983). The authors noted that neither timing of overlaps onset nor their rhetorical function (agreement/disagreement) are reliable predictors of overlap competitiveness. Instead, they claimed that speakers systematically mark competitive overlaps with increased pitch and loudness. These findings were subsequently called into question by Jefferson (1983), who argued that increased pitch and intensity are employed in relatively few cases, indicating that in most cases overlap is resolved by other means. Even

more importantly, she noted that a competitive incomings do not necessarily result in any modification or termination of the overlapped party's speech. The pitch/loudness marking observed by French and Local (1983) might thus have been a result of an unintentional selection bias. Overlap resolution resources, Jefferson (1983) suggested, are likely to be highly context-dependent and not generalisable into a simple decontextualised rule (cf. Jefferson 1984).

Despite Jefferson's (1983) reservations, amplitude and pitch have been repeatedly found to be related to overlap outcomes. For instance, Meltzer et al. (1971) identified a linear relationship between overlap outcomes and amplitude modification of the overlapped party's speech. Importantly, the results suggested that amplitude changes to the incoming party's speech were not related to abandoning the turn by either speaker. The results were further validated by Morris (1971) in an experiment in which intensity of participants' overlapped speech was modified by a computer in real time. It was also found that speakers who are routinely caused to drop out of overlap react not by initiating more overlaps themselves but rather by defending their floor rights more fiercely.

Opposite interaction between F_0 and overlappee/overlapper roles was reported by Bertrand and Espesser (2000), who analysed durational and prosodic features of both the overlapped speech regions and their contexts to predict speaker changes. They found a significant effect of pitch in overlapper's speech and no significant contribution of that parameter in overlappee's speech.

Similarly, Yang (1996) found an association between interruptions and an increase in F_0 and intensity. Furthermore, he argued that the collaborative/competitive distinction forms a continuum with intrusiveness of individual instances of overlapping speech being determined by speakers' urgency to take the turn, as well as their emotional and cognitive states. Therefore, the expected pattern, is a gradual decline in intrusiveness of overlaps as a topic develops due to speakers' information and expression needs being increasingly satisfied and uncertainty levels being reduced.

Wells and Macfarlane (1998) extended on French and Local's (1983) findings by investigating dependencies between competitiveness and overlap onset position within an ongoing turn. They concluded that overlaps are only heard as competitive if, in addition to the pitch and loudness marking, they occur prior to the last major accent in the previous speaker's turn. The authors described two kinds of TRP-projecting accents in British English and redefined a TRP as a stretch of talk between a TRP-projecting accent and turn offset. Similar results were obtained by Cooper (2011).

Finally, Yang and Heeman (2010) reported higher in-overlap mean intensity values for overlap winners than for yielders. Analogous results were obtained for ratios of winners' and yielders' intensity values. Moreover, the ratios were

negatively correlated to overlap duration, meaning that overlaps with smaller intensity differences take longer to resolve. They also found pitch to be a weaker predictor of overlap outcomes than intensity.

With the exception of Bertrand and Espesser (2000), all studies discussed so far focused on characteristics of overlapping speech itself. However, the challenge of such efforts lies in potential feature extraction errors caused by crosstalk (leakage of sound between audio channels). Faced with similar difficulties, some studies have attempted to investigate prosodic features of speech preceding and following the overlapped stretches. Heldner et al. (2010) investigated F_0 contours preceding gaps and overlaps in Swedish maptask dialogues. While their study found some systematic tendencies in prosodic profiles around silences, no conclusive results were obtained for overlaps. By contrast, Gravano and Hirschberg (2012) compared a larger number of context features of overlaps and interruptions. They found that speech preceding interruptions (both involving overlapping speech and pause interruptions) is similar to that preceding within-speaker pauses but with notable differences (e.g. lower mean F_0, faster speaking rate). Previously, Gravano and Hirschberg (2011) analysed overlapped speaker changes (in which the previous speaker was able to complete his or her utterance) and found they were preceded by the same turn-yielding cues as speaker changes accompanied by silence suggesting that those cues are also present earlier in the turn.

More recently, individual contributions of a number of prosodic, temporal and positional features to discrimination of collaborative and competitive overlaps were assessed in a comprehensive study by Kurtić et al. (2013) using a combination of CA and machine learning methods. Their results indicate that while prosodic features (especially a combination of pitch- and intensity-derived features) contribute to overlap classification, features related to position of overlap onset and turn completeness are much more robust. In particular, presence of disfluencies is the most prominent indicator of turn competition. Perhaps more interestingly, Kurtić et al.'s (2013) findings show that, contrary to wide-spread beliefs, there is no deterministic link between position of overlap onset with respect to a TRP and its competitiveness. In other words, both pre- and post-TRP overlaps can be used to compete for the floor, and, similarly, both kinds can be used in a non-disruptive fashion. To complicate the matter even further, different resources were found to be employed in overlaps occurring prior to and within a TRP.

The comparison between the early findings of French and Local (1983) with their single pitch/intensity rule and the manifold interdependencies between features reported by Kurtić et al. (2013) makes it apparent that interruption and turn-competition is indeed a complex phenomenon which has successfully

evaded generalisation and reduction to simple operationalisations. We now turn to another controversial topic in overlap and turn-taking research, namely to the question of how frequent overlaps really are and how long they usually last.

2.3.3 Distributions of overlaps

Sacks et al. (1974) famously postulated that dialogue partners aim at minimising occurrence of gaps and overlaps. As the issue pertains to many fundamental questions such as precision of turn-taking, the presumed target aimed at by speakers and the reactive or predictive mechanisms used for end-of-turn detection, it has stirred many a heated debate. Curiously however, when Sacks et al.'s (1974) claim is put to empirical test, definite conclusions are often difficult to draw. Specifically, Heldner and Edlund (2010) note considerable variability in the proportion of overlaps to all between speaker intervals reported in literature. These span the range from as little as 5% (Levinson, 1983) to over a half of all speaker changes (ten Bosch et al., 2005). A particularly striking example of pervasiveness of overlapping speech was provided by Campbell (2007), who found that speakers in a corpus of Japanese telephone conversations spent more time overlapping another party's speech than speaking on their own. Likewise, diverging results have been reported for average overlap duration. Unfortunately, as pointed out by Heldner and Edlund (2010), the existing studies are only partly comparable due to different measures of central tendency and data transformations used. Specifically, central tendency is commonly expressed by median, arithmetic mean or geometric mean, durations can be log-transformed, and gaps and overlaps can be either analysed separately or treated as part of a single distribution with negative values for durations of overlaps and positive values for gaps.

In Heldner and Edlund's (2010) analysis of turn-taking patterns in Dutch, Scottish English and Swedish corpora of face-to-face, telephone, spontaneous and task-oriented dialogues overlaps longer that 10 ms made up consistently around 40% of all between speaker-intervals in all data sets[4]. Contrary to Sacks et al.'s (1974) predictions only a minority of between-speaker intervals (about 20%) could be classified as corresponding to imperceptible gap or overlap. Rather, the distributions of between-speaker intervals were systematically right-skewed (with modes of about 200 ms), and the most frequent category was a perceptible gap. The proportions remained largely unchanged when recently established perceptual thresholds on overlap detection (estimated at 120 ms) were applied to the data by Heldner (2011). The large observed proportions of gaps and overlaps above detection thresholds led Heldner and Edlund (2010)

[4]Heldner and Edlund (2010) did not analyse within-speaker overlaps.

to conclude that next speakers do not aim at perfect alignment of turns across speaker changes. At the same time, their results were consistent with both predictive (Sacks et al., 1974) and reactive (Duncan and Fiske, 1977) accounts of turn-taking mechanisms.

Less work has focused on overlapping speech in multiparty dialogues. Shriberg et al. (2001) found comparable overlap rates (quantified as percentage of overlapped words and talkspurts) in corpora of telephone conversations and multiparty meetings. However, as expected, they did observe differing results depending on meeting *type* (moderated/non-moderated speaker selection). Çetin and Shriberg (2006) used 26 meetings from various meeting corpora and reported that dialogue participants spend on average 12% of their speaking time in overlap, and 30-50% of talkspurts coincided with background talk. In line with Schegloff (2000), a great majority of overlaps (over 90% in all but one corpus) consisted in simultaneous speech of two parties. Backchannels, floor grabbers and disruptions were found to be more frequently involved in overlap (compared to their overall frequency) than utterances with propositional content, reflecting a systematic link between overlap and floor management[5]. Stretches of overlapping speech were also related to hot-spots (regions of speakers' increased engagement). In addition, Laskowski et al. (2012) found a link between overlap durations and the minimal number of overlapping speakers: overlaps with at least two speakers take shorter to resolve than overlaps with at least three speakers, etc.

2.3.4 Cross-cultural, cross-gender and contextual variation

The no-gap-no-overlap principle has often been challenged from a cross-cultural perspective. It has been claimed that while the system of Sacks et al. (1974) might be a reasonably adequate description of turn-taking in some varieties of American English, it fails to do justice to differences in temporal organisation of speaker selection observed in other cultures. A wide selection of examples has been reported ranging from the North Americans from the west and east coasts to the Thais to the Antiguans and the Puerto Ricans (for a short review see Makri-Tsilipakou, 1994, p. 403). For example, Tannen (2005) described conversations between Californians as characterised by longer gaps and fewer overlaps than found in interactions between New Yorkers. In a similar spirit, Murata (1994) found more interruptions in English than in Japanese conversations. The results

[5] Along similar lines, ten Bosch et al. (2004) suggested that "the group of back-channels [...] obeys other [timing] rules than the rules that speakers adhere to when producing full-content propositions. This result is also compatible with Heldner et al. (2011) as far as their VSUs (very short utterances) are assumed to correspond to backchannels and non-VSUs to propositional content utterances.

was taken to reflect the cooperative character of Japanese interaction. By contrast, as mentioned above, Heldner and Edlund (2010) obtained almost identical results for conversations in Swedish, Scottish English and Dutch, differences in dialogue type notwithstanding. At the same time, their results were largely inconsistent with the no-gap-no-overlap directive.

In an effort to validate claims about inter-cultural variation quantitatively, Stivers et al. (2009) compared timing patterns of responses to polar questions in ten languages (Danish, Italian, Dutch, Tzeltal, ǂĀkhoe Hai‖om, English, Yèlî-Dnye, Japanese, Lao and Korean). They reported similar distributions of gap and overlap durations across all data sets with modes between 0 and 200 ms. However, a more detailed analysis of their results is complicated by the bin size used for data visualisation and the fact that the authors labelled gap and overlap with respect to onset of verbal *or gestural* responses (whichever occurred first). Some of the reported descriptive statistics indicate substantial variability across the analysed data sets (medians ranging from 0 ms to 300 ms, means ranging from 7 ms to 469 ms). Nevertheless, the authors argued that the deviations are small enough to support universality of tendencies towards minimisation of gap and overlap. These universal tendencies were, however, claimed to be mediated by "cultural 'calibration' of delay" (p. 10590), resulting in differences as to what counts as an *interactionally* (rather than perceptually) significant gap and overlap. The calibration was hypothesised to be linked to culture-specific "interactional pace" or "the overall tempo of social life." In other words, even though speakers are claimed to aim at the universally prescribed ideal of the no-gap-no-overlap turn transition, culture-specific factors decide whether or not a gap or an overlap of a certain duration is interactionally problematic. Subsequently, Heldner (2011) applied perception thresholds on gap and overlap detection to Stivers et al.'s (2009) data and observed results similar to those of Heldner and Edlund (2010) with gaps being the most common turn configuration, followed by overlaps and the no-gap-no-overlap category. He also noted similarities among some of the analysed languages (e.g. the Germanic group) and considerable variation among others.

In addition to cross-cultural arguments, spread in literature is a multitude of findings on contextual factors conditioning frequency and duration of overlapping speech. For instance, fewer overlaps were found in conversations between strangers than between relatives (Yuan et al., 2007), in interviews than spontaneous conversations (Jaffe and Feldstein, 1970) and in face-to-face than in telephone dialogues (ten Bosch et al., 2004). However, the problem lacks a systematic and unified treatment. Indeed, conflicting results have also been reported (e.g. lack of differences between relatives and strangers, Shriberg et al. 2001) suggesting that the variability found might be characteristic of specific

corpora with a complicated interplay of various not easily separable factors. Similar problems pertain to issues of gender differences in interruption rates. Over the years this field has accumulated a large body of strikingly contradictory research and is particularly instructive of the confusions arising from attributing differences in overlap rates to a single variable.

As pointed out in Section 2.3.1, the difficulty of studying interruption across categories such as gender consists in a possible interaction between the independent variables (male/female) and recognisability of interruption. Such problems notwithstanding, this line or research has enjoyed unwavering popularity. Zimmerman and West (1975) were probably the first to report an increased likelihood of male interruption into female turn space, a finding based on a very small sample size of about 50 interruptions and interpreted in terms of dominance relations and power struggle in cross-gender interaction. The question has been since revisited by a number of authors, some of whom, like Murray and Covelli (1988), reported opposite patterns using the same coding scheme but a much larger corpus of 400 interruptions. Importantly, they found that a dialogue context as a whole contributes significantly to discriminating between overlaps and interruptions. It should be noted, however, that the interpretation of this finding is complicated by the fact that some of Murray and Covelli's (1988) contexts were *defined in terms of gender*, e.g. a male interviewing a male, a male interviewing a female, etc. Still others (Yuan et al., 2007) found that women are interrupted more than men by interlocutors of either sex. More complex interaction between overlap and speakers' gender were identified by Makri-Tsilipakou (1994) in Greek dialogues. Even though rates of simultaneous speech initiated by men and women were comparable, overlaps fulfilled different roles across genders. While men used it equally frequently to express agreement and disagreement, women used it mostly for affiliative purposes. At the same time, men were more likely to initiate overlaps during women's turns, and women were less likely to overlap their female partners with disaffiliative purposes. In addition, amounts of overlap have been also claimed to differ across same-sex conversations. For instance, ten Bosch et al. (2004) reports higher overlap rates in male-male than in female-female conversations.

The association between interruption and gender is, therefore, far from straightforward. A careful reader is inclined to suspect that the underlying dominance patterns are more complicated than what some authors in the field would like us to believe. While it might be tempting to count instances of overlap initiated by men and women, such a procedure is likely to disregard numerous accompanying factors shaping conversational dynamics. In addition, as shown by Beňuš et al. (2011) (see Section 2.3.7) tracking power relationships between speakers is by no means necessarily tied to employing gender categories.

2.3.5 Multimodal aspects of overlap

Although gestures have been proposed to play an important role in overlap resolution, the issue is largely under-studied. Performance of the widely used intensity-based features for classification of collaborative and competitive overlaps was compared to that of gestural features by Lee et al. (2008). They noted that while increased intensity is effective in correctly identifying competitive overlaps, it performs less well for collaborative overlaps. The opposite was true for gestural features, which can be used to successfully spot collaborative overlaps (characterised by little movement) but less so for competitive ones. As a result, the two feature sets are partly independent markers of intrusiveness and their combination results in a 14% improvement over the intensity-only classification baseline. Subsequently, Lee and Narayanan (2010) used a combination of overlappee's multimodal cues and overlapper's acoustic cues up to one second before overlap onset to predict occurrences of interruption (i.e. non-collaborative overlaps). Features related to mouth, eyebrows and head movements were used. The authors demonstrated that inclusion of overlappee's pre-overlap multimodal features alongside overlapper's acoustic features improves prediction performance. In fact, even when using overlappee's features only, an improvement over chance was observed.

Monada and Oloff (2011) analysed gestures accompanying simultaneous talk in a casual dinner conversation. They identified four gestural strategies employed in overlap resolution: unperturbed gesturing throughout an overlapped region, continuous but perturbed gesticulation, gestural hold followed by continuation and gestural hold resulting in the speaker abandoning the gesture. The four classes were indicative of increasingly problematic overlap instances; the first three were associated with retaining the turn and the last one with abandoning it. Notably, even though Monada and Oloff (2011) considered verbal overlaps only, they rightly point out that gestures preceding talk or not accompanied by talk at all can also be interactionally salient or even interruptive.

While gaze patterns in turn-taking have been studied rather extensively (see Kleinke 1986 for a an overview), overlapping speech has been absent from most studies. Relationships between gazing behaviour and overlap were studied more systematically by Oertel et al. (2012). They concluded that, compared to speaker changes accompanied by silence, overlaps resulting in a speaker change are characterised by a higher likelihood of previous speaker's partner-oriented gaze and a decreased likelihood of the incoming speaker looking away. Less clear patterns were observed for overlaps without speaker change but similar tendencies were discernible. Finally, similar to backchannels produced during a pause and in line with Bavelas et al. (2002), overlaps involving backchannels

showed a clear pattern of mutual gaze extending up to 2 seconds before the overlap onset. Again, the proportion of incoming speakers averting their gaze was higher in overlapped than in non-overlapped positions.

Finally, Steininger et al. (2001) addressed the question of gestures in dialogue system users' barge-ins. Users' in-overlap gestures produced during a series of Wizard-of-Oz experiments were classified into three classes: *interactional* (performed in direct interaction with the system, e.g. pointing), *supporting* (preparatory gestures preceding the next user request, e.g. produced while tracing system output) and a broad *residual* category, comprising emotional and unidentifiable gestures. Overall, they found that many user-initiated overlaps are verbal only. When accompanied by gesture, emotional and unidentifiable gestures were the most common, followed by supporting and interactional gestures. The authors conjectured that the observed patterns might be systematic, as supporting and interactional gestures are likely to be indicative of smooth user-system interactions. By contrast, formally less well-defined gestures might signal interaction problems and could be potentially used for predicting user barge-ins. The authors provided a preliminary characterisation of gestures related to backchanneling, expressing dissatisfaction with system performance and modification of the original request. However, they suggested that, due to high structural variability, the gestures might be better classified using dynamic features (such as velocity).

2.3.6 Overlap in human-computer interaction

In spite of the recent advances in understanding pervasiveness of overlapping speech in human communication, models of turn-taking implemented in dialogue systems perpetuate the emphasis on "clean" speaker changes and the dispreference for user barge-ins. Moreover, most existing dialogue systems use thresholds on silence durations to locate turn ends. In such systems an end of turn is registered if the user has remained silent longer than some predefined period of time (commonly 500–2000 ms). While these efforts are understandable given the impaired ASR performance in regions of overlapped speech (Çetin and Shriberg, 2006), Edlund and Heldner (2005) demonstrated that systems using exclusively silence thresholds are bound to interrupt longer pauses *within* user's unfinished utterances.

At the same time, evidence exists which suggests that barge-in functionality is a highly desirable feature. For instance, Heins et al. (1997) used a simple menu-driven interface to demonstrate that users are likely to spontaneously interrupt the system, and, on discovering that barge-ins are supported, use it

systematically to facilitate interaction.[6] In more complex systems, a dependence between ASR performance and barge-in rates has been reported: only users whose speech was recognised with sufficient accuracy were observed to *deliberately* start speaking during system output (Komatani et al., 2007). Somewhat paradoxically, therefore, unwanted as they are, user barge-in could be considered an indication of system's human-likeness and used to predict ASR errors (Komatani et al., 2008).

As pointed out by Edlund and Hjalmarsson (2012), most systems allowing user barge-in by default terminate their output immediately on detecting simultaneous speech. However, the authors observe that speaker change occurs in less 60% of overlaps found in human conversations, and argue that instantaneous termination of output is likely to be too abrupt a reaction on many occasions. While some strategies have been proposed to prevent a user from taking the floor by temporarily increasing loudness of system output on detecting simultaneous speech, or to delay cessation of system output until real user barge-ins have been successfully distinguished from background noise (Ström and Seneff, 2000) or until backchannels have been distinguished from propositional-content utterances (Edlund et al., 2010), the solutions lack the interactive and incremental character of overlap resolution of human dialogue. The turn-bidding model (see Section 2.1.4) is a particularly attractive proposition, both for dealing with user barge-ins and initiating system barge-ins but has so far only been applied to non-overlapping turn transitions. Another framework for modelling system reactions to user barge-in is a cost-based approach, in which the system action incurring the lowest cost given the current dialogue state is selected. In Raux and Eskenazi's (2009) Finite-State Turn-Taking Machine, for example, a fixed cost is associated with grabbing the floor by the system and an increasing cost is associated with remaining in overlap. A related formulation was proposed by Edlund and Hjalmarsson (2012). In their approach the cost of abandoning the utterance before completion is compared to the cost of speaking in overlap. The cost of incompletion is determined by the information lost should the utterance be terminated prematurely, and the portion produced so far. Consequently, it becomes expensive to abandon nearly finished utterances. By contrast, the cost of expected overlap is related to the inherent cost of producing speech and the additional cognitive load inflicted by speaking in overlap. The relative weighing of the two costs can be performed continuously and incrementally throughout the duration of system output.

[6] Admittedly, the menu-driven system with its repetitive long prompts used in Heins et al.'s (1997) study might have exaggerated the extent of this effect.

2.3.7 Timing of overlap onsets

The bulk of work on overlap timing was done within the field of CA, particularly by Jefferson (1984), who classified overlaps into three categories with respect to their position in a turn: *transitional*, *recognitional* and *progressional*. Overlaps of the first type are for the most part by-products of turn management around TRPs. They consists in talk being initiated in the vicinity of a possible TRP, and the previous speaker continuing his or her turn beyond a completion point. These overlaps will, therefore, coincide with syntactically complete utterances of an interlocutor. Depending on location of overlap onset with respect to completion points in ongoing speech, Jefferson (1984) divided transitional overlap into several classes:

Terminal overlap, in which the next speaker overlaps the very end of previous speaker's turn. Terminal overlaps are, by definition, very brief; however, unlike other types of transitional overlaps, they are not strictly speaking mere by-products of a TRP but are properly projected to intrude into interlocutor's turn.

Terminal Overlap // Overlapped, in which the next speaker starts as above but his or her interlocutor goes on with the turn. Importantly, Jefferson (1984) held that the way such overlaps come about is not in itself a breach of turn-taking rules as both speakers have legitimate rights to the conversational floor.

'Latched'-to-Possible-Completion Onset, in which the onset of simultaneous talk coincides precisely with the completion point of the constituent in progress in interlocutor's speech.

'Unmarked Next Position' Onset, in which the onset of overlapping talk slightly *follows* a TRP. Consequently, the next speaker starts speaking *after* the dialogue partner has embarked on the continuation of his or her turn. According to Jefferson (1984), it is for this reason that such overlaps tend to be perceived as intrusive. However, she argued that overlaps initiated in this position might result from the next speaker starting "in a *blind spot*", not having perceived the interlocutor's continuation.

'Unmarked Next' Position // Overlapped, in which the next speaker starts as above but might in fact have heard the interlocutor's continuation. As a result, even though these overlaps are properly interruptive, their location at "blind" spots helps to legitimise their occurrence.

Two additional transitional overlaps types, *recognitional terminal overlap* and *pre-completor onset*, consist in starting to speak immediately on recognition of how the current speaker is going to complete his or her turn.

The second category of overlap onset, recognitional onset, is related to the previous two categories in that it also depends on recognition of what the interlocutor is going to say. However, recognitional onsets are initiated with respect to speech content rather than turn completion. They do not occur in the vicinity of a TRP and are purposefully designed to coincide with the turn in-progress. For that reason recognitional overlaps can (but do not need to) become interactionally problematic. Finally, progressional onset is a reaction to production problems and disfluencies (silent and filled pauses, stuttering, etc.) in interlocutor's speech. Jefferson (1984) argued that these three overlap categories exhaust the possible locations of simultaneous speech onset within the ongoing turn and provide a systematic account for overlap occurrence regardless of its position.

Importantly, Laskowski et al. (2012) observed that most descriptions of overlap have been framed in terms of "time-independent prior probabilities of its occurrence." While such an approach is informative regarding how likely overlapping speech is overall, it says little about the exact point of its occurrence. Consequently, the authors proposed a dynamic representation of speaker change patterns by means of a Markov process whose states corresponded to the number of speakers vocalising in a 100 ms time window. It was found that such representations are not symmetric due to simultaneous turn starts being more likely than simultaneous turn ends. The difference was largely due to syntactically incomplete or non-propositional utterances. As a consequence, the difference can be successfully used to discriminate between the original and time-reversed conversations. A dynamic perspective was also adopted by Beňuš et al. (2011), who analysed the relationship between temporal patterns of speaker changes in three task-oriented dialogues and establishment of common ground as well as development of power relations among speakers. By combining the CA and quantitative methods they presented some evidence that overlaps can be used to exert time pressure on the interlocutor, and hence constitute an important resource for negotiating dominance structure within interaction and shaping interlocutor's turn-taking strategies. They also demonstrated that entrainment to dialogue partner's speech rhythm is systematically linked to lower dominance. Importantly, Beňuš et al.'s (2011) results indicate that, in sharp contrast to the literature reviewed in Section 2.3.4, studying speaker dominance in dialogue does not require indexing speakers with gender or status categories (in fact, all dialogues used by Beňuš 2011 were balanced for these factors) but can instead rely solely on factors internal to the interaction

itself. On the downside, however, as the authors point out, the study did not use independent measures of speakers' dominance. The authors argue that the observed coordination is best interpreted in terms of affordances offered by the rhythmic organisation of speech. They thus appeal to the regulatory role of rhythm-mediated entrainment between interlocutors (see Section 1.3.5) in speaker transitions. This issue will be discussed in greater detail in the following chapter.

2.4 Conclusions

The above review of literature has attempted to demonstrate the considerable body of findings that has been amassed around the topic of overlapping speech over the years. Overlaps have been studied from many perspectives (discourse analysis, cognitive science, psycholinguistics, phonetics, sociolinguistics) and using (or, indeed, combining) various methodologies (probabilistic modelling, conversation analysis, quantitative corpus-based studies, designed experiments). In fact, the problem of overlapping speech is a clear demonstration of interdisciplinarity of present day investigations into the nature of language, speech and interaction.

Traditionally regarded as a violation of turn-taking laws and a marginal phenomenon, overlapping speech has proved to be both frequent in spontaneous conversation and instructive for understanding organisation and co-construction of human dialogue. In the following chapter we propose that overlapping speech can provide valuable insights into the long disputed role of rhythm and interspeaker entrainment in managing speaker transitions. Specifically, it will be argued that overlapping speech is of particular significance for testing these models by eliminating problematic assumptions and cumbersome measurement methods.

Chapter 3

Models of timing in turn-taking

End-of-turn prediction has been demonstrated to be guided by many different properties of speech. Semantic and syntactic features have been found to be particularly robust indicators of turn offsets (de Ruiter et al., 2006) but prosodic features have also proved to facilitate predicting speaker changes (Ferrer et al., 2002; Gravano and Hirschberg, 2011; Edlund and Heldner, 2005). Somewhat less frequently, it has been suggested that the temporal organisation of turn taking results from speakers becoming entrained to each other's speech rhythm along the lines outlined in Section 1.3.5. Presented there was a model of interpersonal coordination as a rhythm-mediated coupling between speakers resulting in emergence of regularities in the temporal domain. In the context of turn-taking, the tendency to minimise gaps and overlaps (Sacks et al., 1974) was reinterpreted as a consequence of speakers' sensitivity to their dialogue partners' speech rhythm. Turns thus become coordinated in time because of the affordances offered by rhythmicity of speech (Beňuš et al., 2011).

Although, as pointed out in Chapter 2, the claims about the uncanny precision of the speaker change mechanism have since been challenged, the rhythm-based models remain an interesting contribution to explaining the exact instant of turn initiation, whether or not accompanied by gap or overlap, and complement the claims about the role of syntactic and prosodic cues to end-of-turn prediction, which dominate the field. Below, we summarise two such attempts at modelling conversational rhythm by Couper-Kuhlen (1993) and Wilson and Wilson (2005). Subsequently, we discuss some problems the models face and consider how inclusion of overlapping speech might help to solve them.

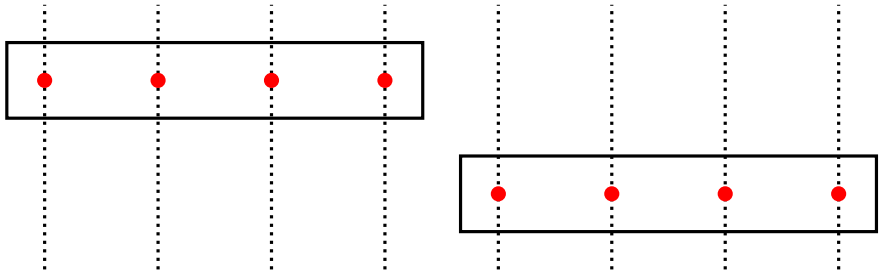

Figure 3.1: Alignment of beats across turns in Couper-Kuhlen's model. Horizontal stripes represent dialogue turns, and circles represent beats.

3.1 Couper-Kuhlen

According to Couper-Kuhlen's (1993) model, the next turn by default continues rhythmic structure of the previous turn, thereby maintaining perceptual isochrony across speaker changes:

> In the unmarked case a next speaker would time his/her entry in such a way that the rhythm and tempo of a prior speaker's talk are maintained. This means (i) that the prior speaker must be speaking in such a way that that perceptual isochrony can be established at some level in the prosodic hierarchy, (ii) that the next speaker monitors enough of this rhythm to ensure a temporally coordinated entry, and (iii) that the first strong syllable of the new speaker's turn comes on the next pulse following the last two rhythmic beats of the prior speaker's turn

Rhythmic beats have been defined elsewhere (Couper-Kuhlen, 1991, p. 275) as "typically created by syllables with relative prominence at some one level of a prosodic hierarchy constituted by syllables, feet, phonological phrases and intonation phrases." The resulting mechanism is presented schematically in Figure 3.1: a turn is well timed if its first strong syllable coincides with the sequence of strong syllables extrapolated from the previous turn.

As Couper-Kuhlen (1993) points out, rhythmisation of speech, which on this view is a prerequisite for accurate timing of the following turn, might be in itself a turn-releasing cue. If rhythmisation is present on multiple levels, an interlocutor is assumed to have a choice of which level to entrain to. In any case, extrapolation of a rhythmic pattern requires at least two strong syllables. Should these be unavailable, "the rhythm must be established over a succession of prior turns" (Couper-Kuhlen, 1993, p. 127), but no account is provided as two how this could be achieved. It is also claimed that in the absence of rhythmic

cues speakers might rely on their spontaneous "natural rhythms." In general, however, "[b]ecause of the gestalt-like nature of rhythmic structures, there is always a 'good continuation,' a next beat at the same rhythm and tempo by the new speaker" (Couper-Kuhlen, 1993, p. 127).

Importantly, one of the consequences of this model is that short silences or overlaps involving weak syllables might be inserted for the sake of maintaining conversational rhythm. By contrast, "real" overlaps involve "temporal coincidence between talk by two speakers which involves rhythmically strong syllables" (Couper-Kuhlen, 1993, p. 131). For the same reason, overlaps are predicted to be more competitive if they disrupt the rhythm of the previous turn:

> There would be two options for 'real' overlap – accommodating to the established rhythm or contravening it. The latter is more noticeable and therefore more marked: it might be thus one way to launch a competitive bid for the floor. (Couper-Kuhlen, 1993, p. 132)

Similarly, silences which disrupt the established speech rhythm or which include "silent beats" (silent multiples of intervals between strong syllables) should be more marked and, therefore, interactionally more significant, for example by indicating potential comprehension problems).

Like most conversation analytic constructs, the model was developed without a backing of a systematic quantitative data analysis. Couper-Kuhlen herself analysed two dialogue transcript and claims to have found support for the model's prediction. Other researchers, however, have been on the whole less successful. In particular, Bull (1996) compared durations of between- and within-speaker intervals in 343 turn exchanges. According to the isochrony hypothesis between-speaker interval durations should be multiples of within-speaker interval durations. This was indeed the case in 40.8% of exchanges. However, given the lack of significant correlation of between- and within-speaker interval durations, the result might have been simply a side-effect of a preference for short intervals (both between- and within-speaker). Szczepek-Reed (2010) studied implications of the model when a speaking turn is transferred from a speaker of a stress-timed language (British English) to a speaker of a syllable-timed language (Singapore English). She observed that in those cases integration of the preceding rhythmic structure is less frequent than reported by Couper-Kuhlen (1993), and is present in only 50% of turn transitions. Additionally, it rarely lasts longer than one stressed syllable. Szczepek-Reed (2010) interpreted this finding as a strategic production of a momentary prosodic link between the turns after which speakers switch to the syllable-timed rhythm imposed by their own language. A notable exception to the general failure of validating the model are studies by Buder and Eriksson (1997, 1999), who used time-series analysis to

identify regular patterns of pitch and intensity peaks across turn transitions with cycles corresponding to inter-stress intervals and breath-groups.

3.2 Wilson and Wilson

In contrast to Couper-Kuhlen's (1993) simple extrapolation-based account, Wilson and Wilson (2005) proposed a model of timing in turn-taking based on the notion of coupled oscillators. The oscillatory character of turn-taking was motivated by regularities in pause durations found previously by Wilson and Zimmerman (1986), who reported that durations of pauses which accompany speaker changes tend to be multiples of a fixed interval ranging from 80 to 180 ms and averaging 120 ms. According to Wilson and Wilson (2005) the periodicity in question might correspond to cyclic application of speaker selection rules such as those put forward in Sacks et al.'s (1974) model, in which the right to self-select as the next speaker alternates between dialogue parties. As a consequence, interlocutors are not equally likely to initiate a turn at all times; rather, the likelihood fluctuates in a periodic oscillatory fashion. As the average period of pause duration corresponds roughly to duration of a single syllable in conversational speech, the frequency of the oscillation might be tied to current speaker's syllable rate (but cf. Cummins, 2012). The function reaches its minimum at syllable midpoints (during sonorous segments such as vowels) and peaks at syllable boundaries (during consonant clusters).

Wilson and Wilson (2005) proposed further that speaker's and listener's oscillators become phase-locked in an anti-phase state (shifted by half a cycle relative to each other, see Figure 3.2). The basis for the mutual entrainment are the cyclic movements of the mandible corresponding to the syllabic organisation of speech (see also Lindblom and MacNeilage 2011). The mutual entrainment is assumed to continue for a couple of cycles into a pause ensuring alternation of speaker's and listener's likelihood maxima and reducing the likelihood of simultaneous starts. However, as a pause gets longer, the oscillators become increasingly decoupled. Notably, the cyclic patterns should be identifiable not only around speaker changes but also in turn continuations by the same speaker.

Regarding the underlying neural mechanism, Wilson and Wilson (2005) hypothesised that the observed cyclicity might be guided by *endogenous oscillators*, groups of neurons found in the human brain which exhibit periodic behaviour and could act as a time frame for other cognitive processes. While their activity covers a wide spectrum of frequencies, the average period of oscillation found by Wilson and Zimmerman (1986) corresponds to the *theta* frequency range (4–10 Hz). Importantly, endogenous oscillators are very robust to signal irregularity and have a strong proclivity for entrainment.

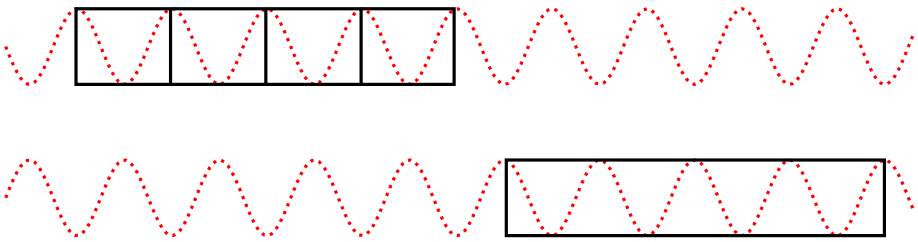

Figure 3.2: Speaker's (top) and listener's (bottom) oscillators in the Wilson and Wilson (2005) model. The oscillators are phase-locked and counter-phased with the frequency of oscillation equal to previous speaker's syllable rate.

An important prediction of this model is that even the most precisely timed turns onsets should not be perfectly aligned with the end of the previous turn. Rather, the phase difference between the oscillators is likely to result in short gaps and overlaps with durations of at least half the syllable cycle. Assuming an underlying syllable rate of 5 syllables per second, gaps and overlaps of 100 ms should be the norm. At the same time, because of detection thresholds on pause and overlap durations (see Heldner 2011), such turns are perceived as perfectly latched onto the previous turn. Additionally, the model predicts matching (or converging) speech rates across interlocutors, and an increase in the number of simultaneous starts in multi-party interactions due to temporal coincidence of maxima in listeners' oscillators. Finally, syllable oscillators could provide a stable reference frame for other temporal phenomena, e.g. phonemic quantity contrasts, through entrainment of low-frequency syllable-related and high-frequency phoneme-related oscillators.

Wilson and Wilson's (2005) model has a number of appealing features—it offers a parsimonious account of timing in turn-taking controlled by low-level and processing-free syllable rate entrainment. It also brings in other advantages of coupled-oscillator approaches: adaptation to variability in speech rate, short response times and compensatory propensity. Oscillators have indeed been demonstrated to be capable of entraining to spontaneous conversational speech (Inden et al., 2012). In addition, there is some evidence that entrainment to low-level features of interlocutor's speech produces more human-like feedback behaviour in embodied conversational agents (Inden et al., 2013). However, the model itself is backed by relatively weak empirical evidence. In particular, Beňuš (2009) tested a number of the model's prediction using both syllables and intervals between consecutive pitch accents as a hypothesised basis for interspeaker entrainment. The data provided limited support for the model. Specifically, weak but statistically significant correlations between speech rates in adjacent turns were found for some utterance types, e.g. continuations af-

ter backchannels (for syllables) or overlaps and pause interruptions (for pitch accents). The correlation for pitch accents improved overall when turns with less than three pitch accents were excluded from the analysis, possibly due to the minimal number of units necessary for successful entrainment. Additionally, pitch accents provided stronger evidence for a relationship between pause length and speech rate in the preceding turn, with the highest correlation values observed for pause interruptions, continuations after backchannels and overlaps. For syllables, the correlation was only significant for pause interruptions. Relative phase of turn onsets was analysed by visual inspection of histograms of ratios of pause duration and syllable rate in the preceding turn. An anti-phase pattern should produce peaks around 0.5, 1.5, 2.5, ... but no such clusters were found. Additionally, the distributions were not visually different from analogous distributions for continuations by the same speaker, where an in-phase pattern (with peaks at 0, 1, 2, ...) was expected. Other predictions of the model concerning isochronous patters in same-speaker continuations, minimal durations of gaps and overlaps and an increased occurrence of simultaneous starts following pauses longer than 1 second were also only partly borne out by the data. Overall, sequences of pitch accents were found to be more conducive to entrainment, especially when limited to certain transition types (interruptions, continuations after backchannels and overlaps).

The Wilson and Wilson (2005) model was further tested by O'Dell et al. (2012) using a single conversation in Finnish. The authors propose a hierarchy of models with increasing complexity, from a baseline of constant pause duration, to Wilson and Zimmerman's (1986) model with pause durations incremented by a constant cycle value, to Wilson and Wilson's (2005) model with variable cycle values dependent on previous speaker's speech rate, to a coupled oscillators model with multiple hierarchical cycles. Neither the constant duration nor the Wilson and Zimmerman's (1986) models provided a strong fit for the data. Significant periodic structure with the mean of 165 ms was only observed for interspeaker pauses of one participant. Moreover, some evidence was found for a secondary turn rhythm with cycles of 0.6-0.8 s, possibly related to units longer than syllables. The results were even less conclusive when, akin to Wilson and Wilson (2005), the cycle length was modelled as dependent on previous speaker's speech rate. As O'Dell et al. (2012) point out, a number of reasons could contribute to such an outcome, the most important of which might be transience of syllable rate:

> Speakers may return to a neutral or preferred turn-taking cycle period fairly rapidly as a pause continues, or natural variation in the period may quickly obscure any initial rate-related difference at the beginning of pause. It may also be that during pauses speakers main-

> tain a turn-taking oscillator for a few cycles only. After all, as pause duration increases, the chance of a simultaneous start decreases even without any synchronisation mechanism. In either case the Wilson & Wilson model will be inadequate, because while allowing turn cycle to vary from pause to pause, it assumes a constant turn cycle during each pause. (O'Dell et al., 2012, p. 225)

The authors also proposed that better estimates of speech rate with weighted history, and using cycles related to other speech units (e.g. feet or morae) could further improve the model's fit. Alternatively, rather than using a single rhythm, a hierarchical model similar to O'Dell et al. (2007, 2008), in which multiple levels of rhythmic organisation are synchronised, could be extended to include pausing.

Subsequently, O'Dell et al. (2012) used hazard regression to model pausing behaviour in dialogue. This technique allows estimating the risk of a participant breaking a pause at any given time providing they have not started speaking so far. Perhaps somewhat surprisingly, the hazard functions were extremely stable across speech rates of previous turn, suggesting little entrainment to rhythmic properties of preceding speech. By contrast, speech tempo had an effect on hazard levels, with faster pre-pausal rates decreasing the risk of other speaker taking the turn and increasing the risk of the same speaker continuing. Additionally, an early maximum prior to 500 ms is present in some speakers (especially for within-speaker pauses), compatible with maximally one syllabic cycle. The resulting bimodal hazard distribution might suggest two parallel processes at work, whereby "a single fast 'turn-taking cycle' is accompanied by a slower background process, with speech starting when one of the processes reaches threshold; relative importance of the two components would apparently vary for different speakers (or situations)" (O'Dell et al., 2012).

3.3 Conclusions

The work outlined in this chapter presented rhythm-based approaches to timing of turn transitions. Two models, by Couper-Kuhlen (1993) and Wilson and Wilson (2005), were discussed. Both assume that orderly timing of speakers' turns is brought about by entrainment to the rhythmic structure of preceding speech, which is continued in the following turn. They are, however, by no means equivalent.

Most significantly, the models differ in their account of interspeaker entrainment. In Couper-Kuhlen's (1993) CA-inspired model, rhythmic congruence is conceived of as an interactional resource, i.e. a conversational strategy speakers

can be seen orienting to. By integrating the rhythmic patterns established by their interlocutor into their own speech, dialogue parties are claimed to produce a prosodic link between turns and demonstrate their sensitivity to sequential organisation of conversation. In other words, presence or absence of turn onsets produced in a timely manner has specific interactional consequences (resulting, for example, in intrusiveness of overlapping speech or in excessively long pauses) for which dialogue participants are accountable and which they address in their later talk.

By contrast, Wilson and Wilson's (2005) model is formulated in terms of physiological and neurological mechanisms determining and regulating fine temporal organisation of turn-taking. Even though deviations from the hypothesised anti-phase pattern are possible, they are likely to result from irregularities in speech production rather than be employed as an interactional strategy. They can, of course, be *reinterpreted* and dealt with as one but the underlying mechanism remains fully automatic. For the same reason, they can only be considered a collaborative achievement up to a point: while rhythmisation of previous speaker's speech should certainly be expected to facilitate entrainment, phase-locking between interlocutors' oscillators is in itself an emergent phenomenon. It should be pointed out, however, that as Couper-Kuhlen (1993) is silent on the point of how the tuning into interlocutor's rhythm is actually achieved, it is difficult to say to what extent she considers the process to be under speakers' active control. Nonetheless, some of the formulation found in the text (e.g. "In the case of rhythmisation at multiple levels, recipients *have the choice* of timing their onsets at different levels.", Couper-Kuhlen 1993, emphasis added) suggest that she does not regard it as a fully automatic and autonomous process.

At the same time, it appears that Couper-Kuhlen's (1993) model requires a certain amount of planning on the part of the next speaker. As she posits entrainment of sequences of strong syllables, a turn needs to be started early (or late) enough and the speech rate of the initial non-accented syllables needs to be suitably fast (or slow) to ensure proper alignment of beats. Unless some automatic compensatory mechanisms between different levels of prosodic hierarchy are assumed, it is difficult to see how these interdependencies could be arrived at without incurring processing costs. Conversely, Wilson and Wilson (2005) propose that the aligned event is the actual speech onset, which is controlled directly by the oscillatory process, thus eliminating the need for explicit speech planning.

The differences notwithstanding, the models tie in with the existing evidence for the role of rhythm in organising human action (see Chapter 1.3.5). However, other than Buder and Eriksson's (1997) and Buder and Eriksson's (1999) results on regularity of pitch and intensity peak sequences across turn transitions,

direct efforts to validate the models using corpus-based methods have met with limited success. We argue that this puzzling fact can be at least partly explained by exclusion of overlapping speech from the analysis.

No account of overlaps is obviously a problem in its own right, since they have been demonstrated to be pervasive in spontaneous dialogue and make up to 40% of all turn exchanges (Heldner and Edlund, 2010). Consequently, ignoring overlaps might result in leaving out of a large portion of data points. However, perhaps somewhat surprisingly given the reputation of overlaps as problematic turn exchanges violating the otherwise smooth speaker transitions, no treatment of overlapping speech might also contribute to the difficulties in finding empirical evidence for entrainment. In fact, presence of pauses between subsequent turns is a serious problem for both models under discussion. Specifically, it necessitates the extrapolation of events in the Couper-Kuhlen (1993) model, and poses the problem of speakers' oscillators becoming decoupled as pause duration increases in the Wilson and Wilson (2005) model . Additionally, in both cases it requires the underlying assumption that the period of oscillation is kept constant throughout the pause. However, this need not be the case. In fact, certain oscillator types, such as McAuley's oscillator (McAuley, 1995), fall back to their eigenfrequency when no input signal is present. More generally, the lack of information exchange serving as the basis for mutual entrainment is a major difficulty in modelling pausing phenomena (O'Dell et al., 2012, p. 226).

Moreover, even though the models make claims about alignment of actual speech events (rather than representational constructs or brain activity patterns), silences between turns make direct methods of inferring entrainment from the data impossible. Again, extrapolation of the preceding rhythmic pattern into a silent interval requires resorting to ratios of pause duration (as in Bull, 1996) or to estimates of speech tempo *over the entire* preceding speech chunk (as in Beňuš, 2009 and O'Dell et al., 2012). Such methods enforce a static view on what is a fundamentally dynamic process, and might, therefore, miss their target entirely. In addition, the averaging involved in these approaches is likely to obliterate any subtle adaptation processes.

By contrast, no such difficulties arise when stretches of overlapping speech are considered. Since in overlapping speech both interlocutors are by definition speaking simultaneously, there is no need for extrapolation of previous rhythmic patterns. Neither should overlapping turn exchanges result in speakers' oscillators becoming decoupled as there is constant bi-directional exchange of information between dialogue partners.[1] Finally, since the rhythm interlocutors

[1] It could perhaps be argued that clashes between speech production and perception operating in parallel during stretches of overlapping speech might indeed weaken the coupling between speakers. However, this issue has no direct import for the present work since we are concerned

entrain to is the actual rhythm produced concurrently by their dialogue partners, there is no need to employ measures of periodicity derived from speakers' past behaviour. Instead, the relative timing of any events in interlocutor's speech can be measured directly from the data. The following section outlines one such method of measuring and analysing timing of overlapping turn onsets, which will be used in the remainder of the work.

with timing of overlapping turn onsets, which are by definition preceded by speech of only one dialogue participant.

Chapter 4

Method

In the previous chapters rhythm has been postulated as an important organising principle behind human behaviour in general, and turn transfer in particular. However, the rhythmic component to turn-taking has so far evaded empirical verification. At the same time, overlapping speech has been demonstrated to be omnipresent in spontaneous interactions but largely missing from earlier studies of temporal entrainment in turn-taking, which, as we argued, has serious theoretical and practical consequences. Therefore, the remainder of this work is devoted to a study of temporal patterns in overlapping speech.

Specifically, given the properties of overlapping speech allowing direct measurement of timing patterns (see Section 3.3), the simple method for capturing interspeaker entrainment described below was employed. In short, the method quantifies the location of simultaneous speech onsets relative to various landmarks (e.g. syllabic boundaries) in interlocutor's turn. If interspeaker synchronisation occurs, systematic deviation from a random baseline should be observed, along the lines outlined in Section 4.2.

Instances of overlapping speech were derived from interpausal units (IPUs), i.e. stretches of speech bounded by at least 100 ms of silence and / or laughter. Timing of overlap onsets was then calculated with respect to various prosodic and linguistic boundaries in the following manner: For each overlap, the first overlapped unit of analysis (*target unit*, e.g. a syllable) of the overlappee's IPU (i.e., the unit during which the overlap was initiated) was identified. The overlap onset was then normalised relative to the duration of this target unit. Specifically, the *normalised overlap onset time* was calculated by dividing the duration of the interval from the onset of the overlapped unit to the onset of the overlapping utterance by the duration of the overlapped unit. The procedure is illustrated in Figure 4.1.

The resulting value represents the overlap onset time as a fraction (with values between 0 and 1) of the target unit duration; that is, the proportion of

the target unit duration after which the overlap was initiated. Equivalently, normalised overlap onset time can be interpreted in terms of phase, i.e. "the relative time of an event with respect to some containing and repeating unit" (Cummins, 2011b, p. 1).

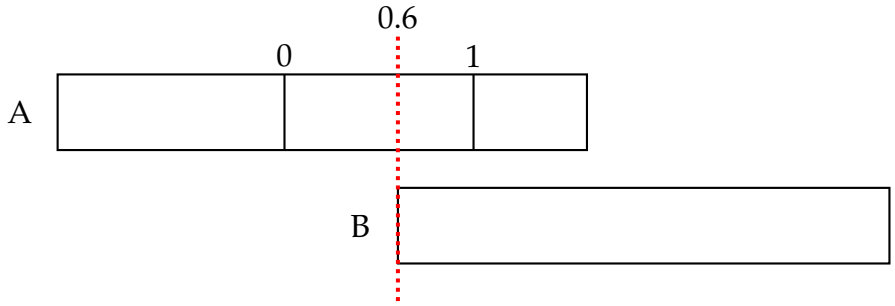

Figure 4.1: Overlap onset relative to the duration of the first coinciding target unit in overlappee's speech. The top stripe represents overlappee's speech, 0 and 1 mark the boundaries of the target unit. The bottom stripe represents the overlapping IPU.

The majority of analyses reported in Chapter 5 were done using Switchboard (Godfrey et al., 1992), a corpus of spontaneous (non-scripted) telephone conversations between strangers on a preassigned topic. To ensure the findings are not corpus- or language-specific, some of the analyses were also repeated on three other corpora:

- the CID corpus (Bertrand et al., 2008, sldr000720): a multimodal corpus of spontaneous face-to-face French conversations,
- the Kiel Corpus of Spontaneous Speech (the "Lindenstraße corpus", IPDS, 2006): a corpus of non-face-to-face (speaker were seated in separate rooms) German conversations between friends about the German TV soap opera *Die Lindenstraße*,
- the Finnish Dialogue Corpus (Lennes, 2009; O'Dell et al., 2008): a corpus of spontaneous Finnish conversations between friends.

For English, German and Finnish IPU boundaries were derived from word segmentations distributed with the corpora. For French, we used the IPU segmentations distributed with the corpus, with the exception of IPUs starting with or consisting entirely of non-verbal phenomena, such as laughter. Those IPUs can, therefore, contain non-initial stretches of laughter, which were treated as equivalent to silence in the other corpora. The somewhat different treatment

of laughter in the French corpus and in the other data sets might have produced a slight asymmetry in IPU segmentations. However, in the light of the consistent results presented below, we feel the differences can be safely ignored.

The following target units in overlappee's speech were used: syllables (in the Switchboard and CID corpora), intervals between consecutive vowel onsets (henceforth, vowel-to-vowel intervals, VTVs; for all four corpora), intervals between consecutive pitch accents (henceforth inter-accent intervals IAI; for a subset of 75 dialogues in the Switchboard corpus) and words (in the Switchboard corpus).

All annotations used were distributed with the corpora. In the Switchboard and CID corpora all boundaries with the exception of pitch accents had been produced by automatic segmentation methods. Segmentations in the Kiel Corpus and the Finnish Dialogue Corpus had been performed manually.

4.1 Central tendency measures

Notably, normalised onset time calculated according to the procedure in Figure 4.1 is by definition a quasi-cyclic measure[1], as the value of 1 in one unit is identical with the value of 0 in the following unit. In other words, 0 and 1 in *contiguous* units represent what is essentially the same point (syllable boundary, lexical stress, pitch accent, etc). The cyclic nature of the data should become even more evident when two hypothetical overlap onsets occurring immediately after and immediately before a target unit boundary are considered. Such overlaps are assigned values from the opposing ends of the scale (0 and 1 respectively) even though they are in fact produced in very close proximity.

For this reason the familiar linear measures of central tendency and dispersion (arithmetic mean, standard deviation, etc.) are inappropriate for analysing data of the kind considered here. Instead, below we define basic descriptive measures following their interpretation in circular statistics, in which each data point is represented as an angle ϕ_i on a unit circle (see Batschelet, 1981).

Central tendency is represented as a vector, whose *length* is calculated from mean rectangular coordinates of individual angles ϕ_i:

$$r = \sqrt{\overline{x}^2 + \overline{y}^2},$$

$$\text{where } \overline{x} = \frac{\sum_{i=1}^{n} \cos \phi_i}{n},\ \overline{y} = \frac{\sum_{i=1}^{n} \sin \phi_i}{n}$$

[1] As we will argue in Chapter 5, the actual observed distributions are produced by at least two independent processes, one of which is circular and one which is not.

The length of the mean vector is a measure of data concentration with more concentrated samples producing higher mean lengths. However, it should be noted that low mean vector length values do not necessarily result from randomly distributed data points. For example, a mean vector of length 0 will be obtained for a bimodal circular distribution with identical peaks π apart.

The average *angle* $\overline{\phi}$ is given as:

$$\overline{\phi} = \begin{cases} \arctan(\overline{y}/\overline{x}) & \text{if } \overline{x} > 0 \\ 180^\circ + \arctan(\overline{y}/\overline{x}) & \text{if } \overline{x} < 0 \\ 90^\circ & \text{if } \overline{x} = 0 \text{ and } \overline{y} > 0 \\ 270^\circ & \text{if } \overline{x} = 0 \text{ and } \overline{y} < 0 \\ \text{undetermined} & \text{if } \overline{x} = 0 \text{ and } \overline{y} = 0 \end{cases}$$

To measure data dispersion *circular variance*, defined as $2(1 - R)$, can be used as an alternative to mean vector length. Finally, *circular range* is defined as the smallest arc enclosing all data points. The measures are summarised graphically in Figure 4.2.

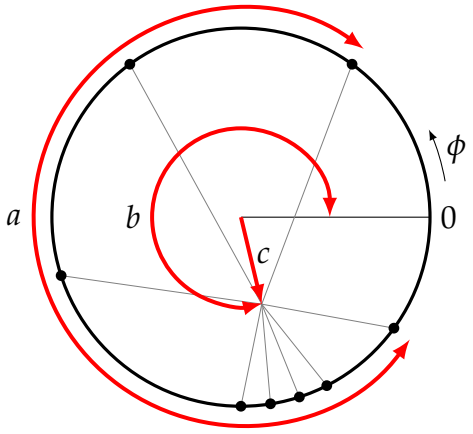

Figure 4.2: Schematic illustration of circular range (a), mean angle (b) and mean vector length (c). Individual angles are represented as points on the circumference.

4.2 Hypotheses

In the light of the models of turn timing and the postulated temporal dependencies between speech onsets and rhythmic sequences in preceding turn outlined

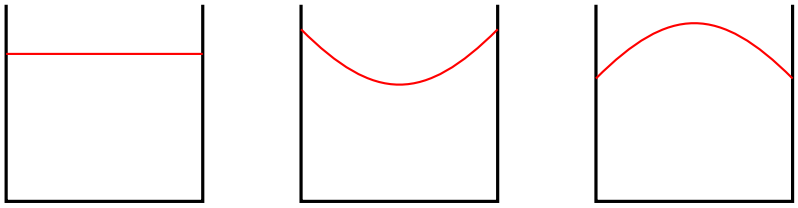

Figure 4.3: Expected distributions of normalised onset times given random timing of overlap onsets (left), an in-phase (middle) and an anti-phase pattern (right) between overlap onsets and target unit boundaries.

in Chapter 3, certain expectations concerning distribution of normalised overlap onset time within target units can be formulated.

The expected patterns are presented in an idealised form in Figure 4.3. More precisely, completely random timing of overlap onsets should produce the uniform distribution plotted on the left. Here, an overlap initiation is equally likely at each point within a target unit. Any deviation from this random pattern must result in systematic clustering of overlap onsets. Specifically, if overlap onsets and target unit boundaries exhibit an in-phase pattern, overlaps are more likely to occur near target unit boundaries and the U-shaped distribution shown in the middle should be expected. By contrast, an anti-phase pattern should result in overlaps occurring more likely near target unit midpoints, producing the peak portrayed in the right panel of Figure 4.3. Other phase values will produce peaks lying in between those extremes.

Distributions of normalised onset times were compared against the uniform (flat) baseline with the one-sample Kolmogorov-Smirnov test (1-KS). Analogously, differences between empirical distributions were tested with the two-sample Kolmogorov-Smirnov test (2-KS). While circular tests of uniformity exist (e.g. Kuiper's test), they were deemed inappropriate given the observed discontinuities around distribution edges, suggesting presence of non-circular influences (see Section 5.5.1). For the same reason histograms were used for data visualisation instead of density estimates using circular kernels.

In addition, deviation of the observed ranges from randomness were tested with a mean range test consisting in repeatedly (N=10000) drawing random samples of the same size as the observed ones from a circular uniform distribution, with circular range calculated in each iteration. *p*-values were subsequently derived as the proportion of random samples with the range at least as extreme as the one observed.

Chapter 5

Results

The present chapter presents results of the study into timing of overlap onsets within a range of target units from different levels of phonetic, linguistic and conversational organisation. Section 5.1 starts with a discussion of overlap onset alignment with respect to syllabic boundaries. The results are subsequently revised in Section 5.2 using intervals between consecutive vowel onsets, which are claimed to provide a more robust basis for entrainment than phonologically defined syllable boundaries. The results are reproduced on data from four typologically different languages, suggesting a universal character of the underlying process driven by prominences in speech. The latter are subsequently studied in some detail in Section 5.3. Section 5.4 investigates effects of preceding context rhythmisation on overlap timing. Finally, Section 5.5 turns towards high-level influences of lexical boundaries, phonological syllable weight, position within an IPU as well as functional influences of utterance type.

5.1 Overlap onsets within syllables

Timing of overlap onsets within syllables is analysed first. Figure 5.1 shows the distribution of syllable-normalised onset time for 10272 overlaps in the Switchboard corpus. Overlaps coinciding with overlappee's IPU-initial and IPU-final syllables were excluded from the analysis with a view to eliminating simultaneous starts and terminal overlaps, which are likely to be related to predicting utterance rather than syllable boundaries. The bimodal shape of the distribution indicates that timing of overlap initiations is not uniform within a syllable. Instead, overlaps are more frequent around syllable boundaries than towards its centre, hinting at an in-phase pattern between overlap onsets and syllabic boundaries (see Figure 4.3). This is confirmed by a statistically significant result of the one-sample Kolmogorov-Smirnov test ($p < 0.001$).

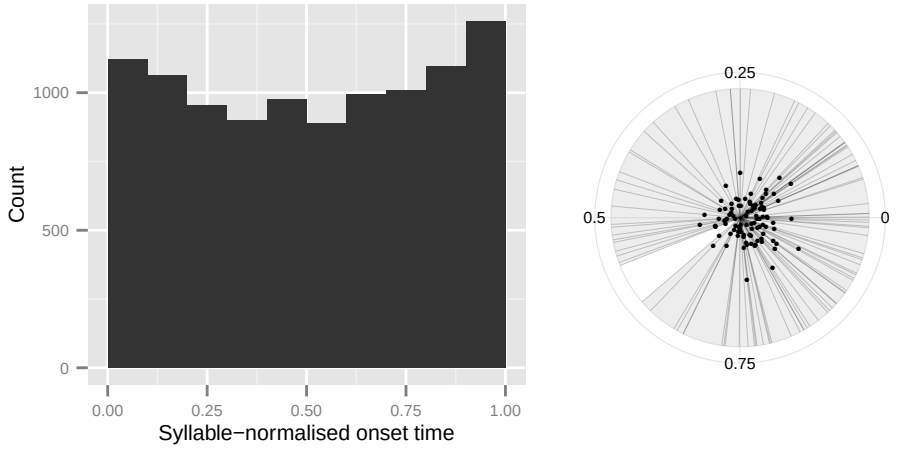

Figure 5.1: Distribution (left) and subject means (right) of syllable-normalised onset time in the Switchboard corpus. Range is represented by the shaded area.

To ensure the result is not an artefact of the employed measurement method, an analogous distribution was obtained for overlaps between speech of randomly paired speakers. Since each speaker was picked by chance, the constructed "dialogues" should also exhibit a completely random pattern. The resulting distribution (Figure 5.2) is indeed neither visually nor statistically ($p = 0.57$, 1-KS) different from a flat uniform distribution. In particular, the distribution does not exhibit the bimodal shape visible in Figure 5.1. The effect is thus independent of data pre-processing and analysis methods.

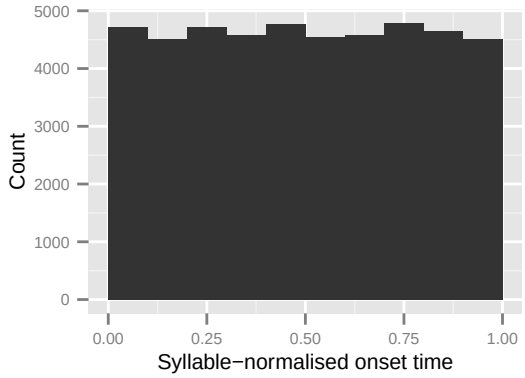

Figure 5.2: Distribution of syllable-normalised overlap onset time for randomly paired speakers in the Switchboard corpus.

The observed non-randomness of overlap timing provides support for inter-speaker entrainment and is compatible with the role of rhythm hypothesised in

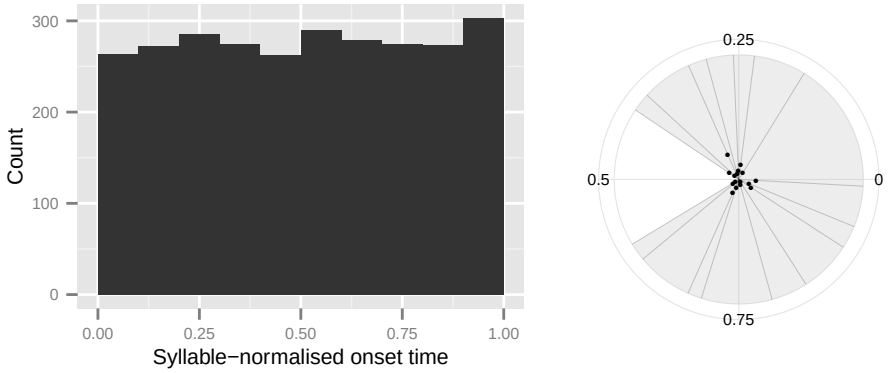

Figure 5.3: Distribution of syllable-normalised overlap onset time in French. The range is represented by the shaded area.

Chapter 3. However, Szczepek-Reed's (2010) results imply that differences in rhythmic properties between languages might influence patterns of interspeaker coordination. Consequently, a question arises concerning a language-specific nature of the effect. To ascertain universality of the underlying mechanism, syllable-normalised overlap onset time was also calculated for the French corpus. The effect, however, was not reproduced: the resulting distribution, shown in Figure 5.3, is visibly flatter than that for English. In addition, the Kolmogorov-Smirnov test did not reveal significant deviation from randomness ($p = 0.48$).

The difference between the results obtained on the English and French data sets might be indicative of language-specific entrainment patterns, in line with Couper-Kuhlen (1993) and Szczepek-Reed (2010). On that view, overlappers are expected to entrain to those levels of rhythmic organisation which are particularly strongly marked in a given language. Specifically, as more regular sequences should be easier to entrain to than irregular ones (see also Section 5.4), rhythmisation on different prosodic levels should lead to diverging entrainment patterns across languages. For instance, it has been long claimed that while some languages are characterised by regular intervals between stressed syllables (stress-timed languages, such as English or German), others show regularities on the syllabic (syllable-timed languages, such as French) or moraic (mora-timed languages, such as Japanese or Finnish) levels. These claims have been repeatedly challenged but in a weaker form remain a workable approximation of *tendencies* towards isochrony within a language. However, even though it is certainly plausible that rhythmic properties of a particular language should favour certain units as a basis for interspeaker entrainment, under these assumptions French should show a *stronger* effect than English as syllabic sequences in French are purportedly more regular and, consequently, easier to entrain to.

For this reason, even though the language-specific nature of the coordination cannot be ruled out completely at this stage, an alternative account presents itself. Namely, the syllable-based pattern observed in English could be a side-effect of entrainment to a different sequence type, which is also involved in speaker synchronisation in French. In other words, even though speakers entrain to the same type of events in both English and French, the effect is only visible within syllabic intervals for English. As this hypothesis naturally favours a view of entrainment grounded in universal properties of speech perception and production, the events allowing synchronisation should also be perceptually motivated.

Among those, *perceptual centres* (p-centres) are particularly suitable for providing a language-independent and perceptually motivated entrainment basis. Crucially, p-centres, which most commonly coincide with vocalic onsets, have been proposed as *perceived* syllable onsets (Morton et al., 1976). There is also evidence suggesting they might be implicated in sensorimotor coordination, for example when tapping to speech (Allen, 1972) or synchronising speech with external rhythms (Rapp-Holmgren, 1971).

Given the perceptual salience of p-centres and their possible role in synchronisation of motor action with external stimuli, English and French syllable-normalised results were recalculated using intervals between consecutive vocalic onsets (vowel-to-vowel intervals, VTV). A similar procedure was also carried out on the Finnish and German data sets. The results are presented in the following section.

5.2 Overlap onsets within intervocalic intervals

Above we hypothesised that the clustering of overlap onsets around syllable boundaries in English might in fact reflect entrainment at the level of VTV intervals. However, as average durations of syllables and VTV intervals are comparable, replacing the one target unit with the other should produce distributions equivalent in shape but shifted in a consistent manner. More precisely, as vowel onsets coincide with or, more frequently, follow syllabic onsets, using the latter should in most cases result in a shift of target unit boundaries to the right (except for syllables starting with a vowel, in which case the boundaries stay intact). The expected pattern is thus a decrease of normalised onset time values, which effectively moves the resulting distribution to the left (or, equivalently, shifts the zero point to the right), producing a late peak. The procedure is illustrated schematically in Figure 5.2.

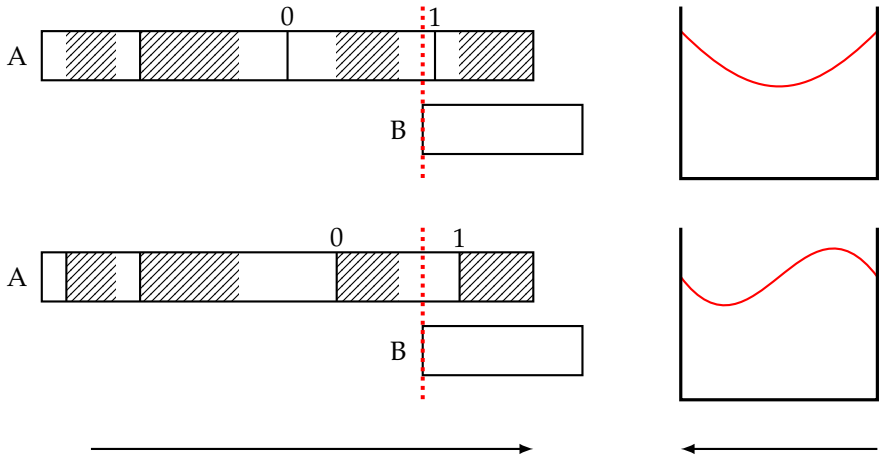

Figure 5.4: The expected shift of the normalised onset time distribution resulting from replacing syllable boundaries (top) with vocalic onsets (bottom). The stripes on the left represent speakers' IPUs, vocalic intervals are marked with shaded areas. 0 and 1 designate target unit boundaries.

Distributions of normalised onset time for the four corpora (with 10858, 2738, 1028 and 353 overlaps for English, French, Finnish and German respectively[1]), plotted in Figure 5.5, follow this expectation. A similar pattern is discernible in all four data sets: the likelihood of overlap initiation is the smallest directly following the vowel onset (0) and peaks around 80% of the VTV duration, i.e. shortly before the upcoming vowel onset (1). The effect is most manifest for English with less clear but compatible patterns in the other languages. In addition, the distributions for both English and French are significantly different from a uniform distribution ($p < 0.001$, 1-KS). While a non-significant result was obtained for German and Finnish ($p = 0.12$ and 0.22 respectively, 1-KS), it is likely to have resulted from substantially smaller sample sizes.

Curiously, the distributions in Figure 5.5 feature strong discontinuities around the edges of the scale, especially in English, German and, to a lesser extent, French, which is at odds with the circular character of the employed measure (see Chapter 4). In Sections 5.3 and 5.5 we claim that the discontinuities stem from non-circular influences of long VTVs and IPU boundaries.

Mean angles for individual languages are plotted in Figure 5.6, and are consistent with a preference for late overlap onset. Not surprisingly given the subtlety of the effect, mean vector lengths are small, suggesting high in-group variances. More striking, however, is the consistency of mean angles, all of which cluster around 75% of the VTV duration, resulting in a very small range.

[1]Overlaps coinciding with IPU-initial VTVs were excluded from the analysis.

Mean angle values for individual speakers are shown in Figure 5.7. Here, again, most overlappers can be seen to start speaking late in the VTV, which is reflected in a higher concentration of data points in the second half of the circle. Mean range test was used to calculate the probability of the observed ranges given that samples were drawn from a uniform distribution. As before, the result is statistically significant for English ($p = 0.02$) and French ($p = 0.016$) but not for Finnish ($p = 0.052$) and German ($p = 0.454$).

Collectively, these results are consistent with the syllable-based results reported for English in the previous section given the expected shift shown in Figure 5.2. In addition, however, a statistically significant result was obtained for French. As no effect was observed on the same data set when syllable boundaries were used, it might suggest that, in line with p-centre literature, vowel onsets provide a more robust and perceptually salient basis for synchronisation than do phonologically defined syllable boundaries. Coupled with comparable results for all four languages, this hints at a perceptually-motivated universal effect. This point is dealt with further in the following section, which investigates links between overlap timing and other perceptually prominent speech events.

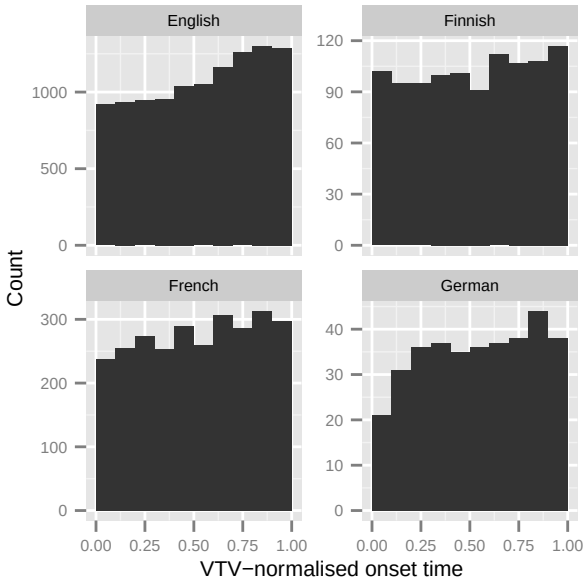

Figure 5.5: Distributions of VTV-normalised overlap onset time in English, Finnish, French and German.

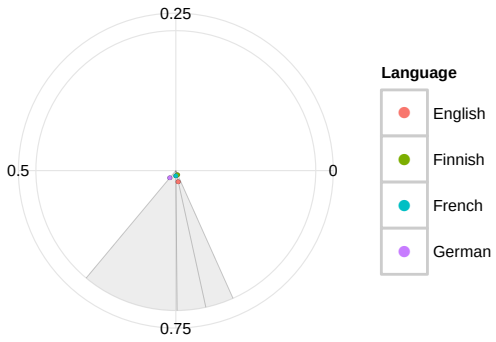

Figure 5.6: Mean VTV-normalised overlap onset time in English, Finnish, French and German. Range is represented by the shaded area.

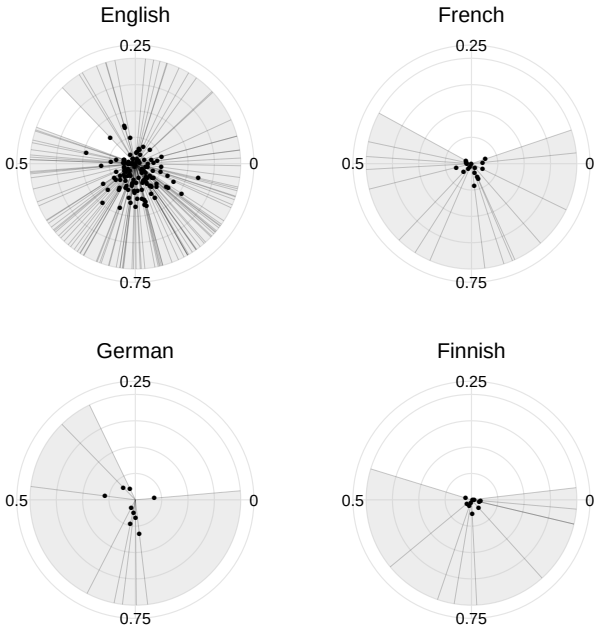

Figure 5.7: Subject means of VTV-normalised onset time in English, Finnish, French and German. Ranges are represented by the shaded area.

5.3 Prominence-related effects: pitch and duration as a basis for entrainment

An extension of the stipulated perceptually motivated nature of the investigated effect relates frequency of overlap onsets to perceptually prominent events in the flow of speech. For this reason the sonorous vocalic segments have been hypothesised to provide a more stable and more perceptually robust basis for entrainment than the relatively arbitrary syllable boundaries commonly coinciding with less prominent consonantal sounds. The present section explores this hypothesis further by evaluating the influence of pitch and duration on temporal patterning of overlapped speech onsets, both of which have been shown to be linked to perceptual prominence in English (Fry, 1958; Silipo and Greenberg, 1999) as well as other languages (Jessen et al., 1995; Portele et al., 2001; Malisz and Wagner, 2012; Fant et al., 2000; Goldman et al., 2007)

In order to assess the impact of accentuation, onsets of 827 overlaps in 75 Switchboard dialogues annotated with pitch accent labels were normalised to the duration of the first overlapped interval between consecutive pitch accents (inter-accent interval, IAI) in overlappee's speech (see Figure 5.8). Overlaps coinciding with VTVs preceding the first and following the last pitch accent in overlappee's IPUs were excluded from the analysis.[2]

The resulting distribution is plotted in the left panel of Figure 5.9. Overlaps are least likely to be initiated around pitch accents with a broad peak around the 60% of the IAI duration. In other words, they are more frequent in the latter half of the IAI. The distribution is significantly different from a uniform baseline ($p < 0.001$, 1-KS).

Notably, given that F_0 changes associated with pitch accents extend over some time domain preceding and following the peak, skewness of this distribution might be interpreted as a measure of the relative strength of preceding and upcoming pitch accents in pushing away overlap initiation. Specifically, the negative skew observed in the present case might indicate that the effect of a preceding pitch accent is stronger than that of an upcoming pitch accent. This is expected insofar as a past event is likely to exert greater influence on (overlapper's) speech perception and production than an event still to be produced.

For comparison, VTV-normalised onset time calculated on the same 75 Switchboard dialogues is presented in the right panel of Figure 5.9. The distribution is significantly different from random ($p < 0.001$, 1-KS) and similar in

[2]It was found out subsequently that in 13 of the 75 analysed dialogues accents were labelled only for a subset of sentential clauses (representing the majority). We leave reassessment of the results using only the fully annotated conversations for future work.

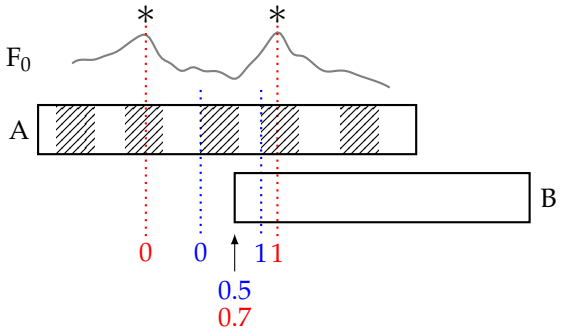

Figure 5.8: Overlap onset relative to the duration of the first coinciding inter-accent interval (red) and vowel-to-vowel interval (blue) in overlappee's speech. The stripes represent speakers' IPUs, vocalic intervals are marked with shaded areas, and pitch accents with stars.

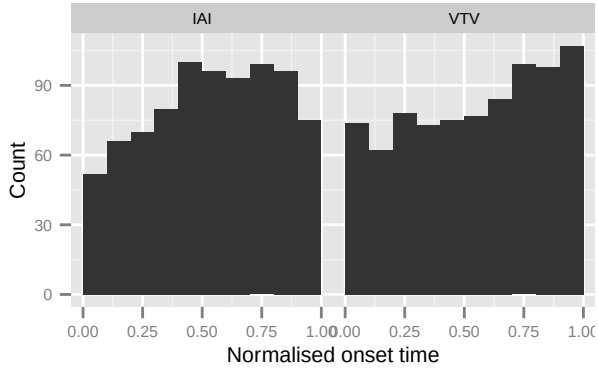

Figure 5.9: Distributions of overlap onset time normalised to the duration of the IAI (left) and VTV (right) coinciding with the overlap onset.

shape to distributions obtained for the whole corpus and the other languages analysed previously (see Figure 5.5).

In the light of the non-random timing of overlap initiation both within VTVs and IAIs, the relation between the two observed effects becomes of interest. Can one be explained in terms of the other? Can they be both traced down to a common cause? Or are they more or less independent of each other? Indeed, given that rhythm in speech, the stipulated basis for the kind of speaker adaptation discussed here, is a hierarchical phenomenon, some of the observed patterns might in fact be side-effects of entrainment occurring at higher levels of phonological hierarchy.

To answer these question and separate the various possible contributing factors, the VTV distribution in Figure 5.9 was split depending on whether

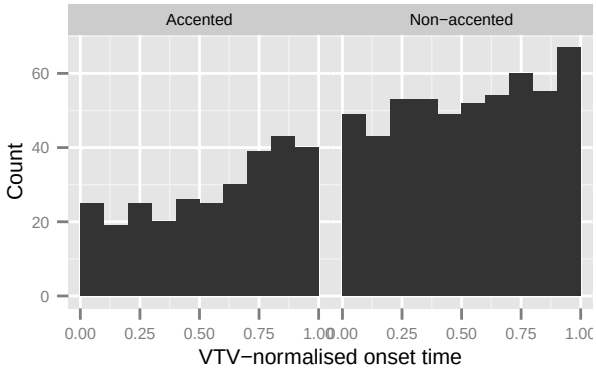

Figure 5.10: Distributions of normalised overlap onset time within accented and non-accented VTVs.

the VTV coinciding with the overlap onset carried pitch accent. If presence of pitch accents was indeed the main cause of the observed effect, non-accented VTVs should follow a random pattern. The resulting distributions, plotted in Figure 5.10, are in line with this hypothesis: the distribution of accented VTVs is markedly less flat than that of non-accented ones, and unlike the latter is significantly different from a uniform distribution ($p < 0.001$ and $p = 0.085$ for accented and non-accented VTVs respectively, 1-KS).

While the above results suggest that presence of F_0 movement has an effect on timing of overlap onsets, lack of a statistically significant outcome for unaccented VTVs certainly does not warrant inferring lack of effect. Additionally, the accented / non-accented dichotomy potentially conflates pitch and duration. With a view to separating individual contributions of each feature, accented and non-accented VTVs were further split on their median durations (291 and 190 ms respectively). The resulting distributions are plotted in Figure 5.11.

Each of the categories in Figure 5.11 was compared against a uniform distribution yielding the following p-values (1-KS): $p < 0.001$ (long accented, long non-accented), 0.95 (short accented), 0.64 (short non-accented). The fact that the distribution of long non-accented VTVs displays a non-random pattern both statistically and visually whereas that of short accented ones does not might indicate that duration is the main (or indeed the sole) factor influencing overlap initiation patterns with little or no contribution of pitch. However, these results need to be taken with caution given the small sample counts, especially in the accented category, and the known subtlety of interspeaker adaptation phenomena. Indeed, while duration might be the main perceptual cue guiding overlap initiation, it could nevertheless be mediated by pitch modulation. More generally, individual contributions of the two features will be

difficult to disentangle because of the inherent lengthening effect of accentuation (but see Tamburini and Caini 2005 for an attempt).

These reservations are confirmed by results in Figure 5.12, in which overlap onset values are plotted separately for long and short VTVs (again split on median duration, equal to 220 ms) in the entire Switchboard corpus. Two things are of import here. First, short VTVs exhibit a non-random pattern ($p < 0.01$, 1-KS). At the same time, the effect is very subtle and, therefore, difficult to ascertain with small sample sizes. Second, in long VTVs overlap likelihood tends to rise towards the end of the unit with no indication of a similar increase near its beginning. This is somewhat surprising since, as we pointed out in Section 4.1, our measure of timing is essentially circular and many of the overlaps produced in the vicinity of vowel onsets should be expected to fall into the next VTV, resulting in comparable frequencies on either end of the scale. This is in fact the case in short VTVs, which suggests that the pattern observed on long VTVs is not merely quantitatively but also qualitatively different, and corresponds to an underlying non-circular process.

To trace the link between overlap onset time and VTV duration in greater detail, in Figure 5.13 we plot a heatmap of normalised onset time in VTVs with durations between 50 and 500 ms. Frequencies were scaled to 1 column-wise, so that the gradient corresponds to relative frequency of onset time values

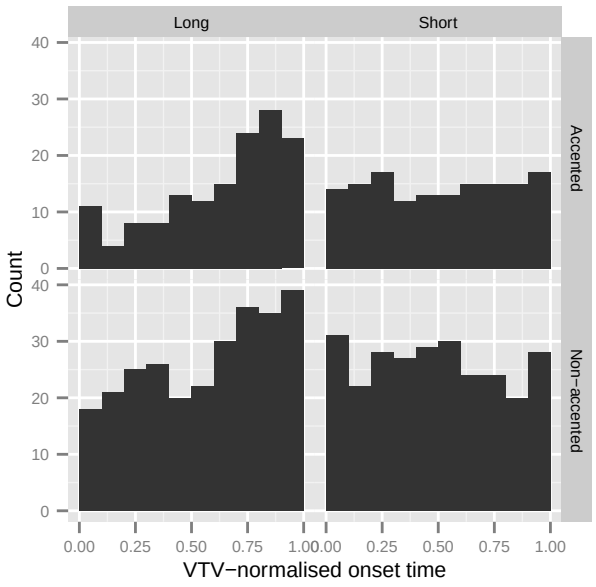

Figure 5.11: Distributions of normalised overlap onset time within accented and non-accented VTVs split on median durations.

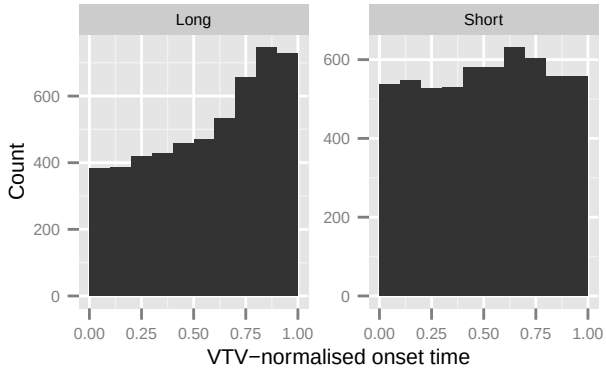

Figure 5.12: Distributions of overlap onsets in short (< 220 ms) and long (> 220 ms) VTV intervals in English.

for a given range of VTV durations. The figure makes clear that the marked tendency for the final rise is indeed characteristic of longer VTVs, evidenced by the high concentration of points in the upper right corner of the plot. In fact, the trend is only present in VTVs with durations of 200 ms upwards; in shorter intervals a weaker clustering of data is discernible between 50 and 80% of the VTV duration.

Similar results are obtained when mean overlap onset time values are calculated for VTVs of increasing durations over a series of overlapping windows of 10 ms with a 5 ms step. The means are plotted in Figure 5.14 against a semi-transparent scatterplot allowing a visual assessment of data concentration. They confirm the conclusions reached above. Notably, a clear discontinuity in mean overlap onset time can be observed coinciding with a VTV duration of about 200 ms, thereby providing a further justification for the data split performed above. More importantly, it emphasises the qualitative difference between temporal alignment of overlap onsets in long and short VTVs. In VTVs shorter than 200 ms mean onset time values drop steadily from around 0.9 towards 0.5. By contrast, overlaps are consistently initiated towards the end of longer VTVs.

Subsequently, a random sample of overlaps coinciding with VTVs longer than 200 ms was inspected manually. It was found that in those cases simultaneous speech is often invited by certain prosodic cues, such as backchannel-preceding cues (Gravano and Hirschberg, 2011). Otherwise, the lengthening itself was a frequent marker of topic structure, phrasing or disfluent speech.[3]

[3]In addition, some of the longer overlaps were found to result from segmentation errors consisting in segments spanning silent pauses. These cases, therefore, are more likely to correspond to simultaneous starts.

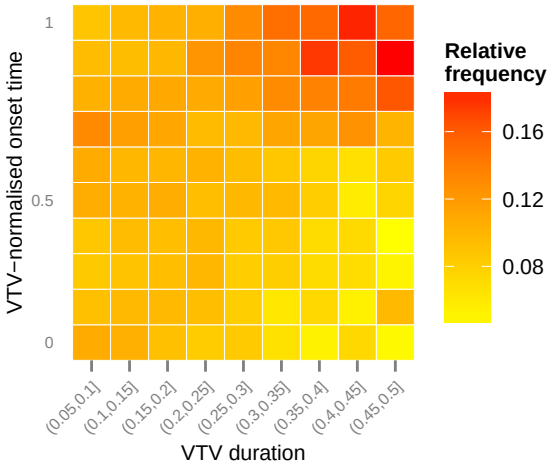

Figure 5.13: A heatmap of the overlapped VTV duration against VTV-normalised onset time. Frequencies were scaled to 1 column-wise.

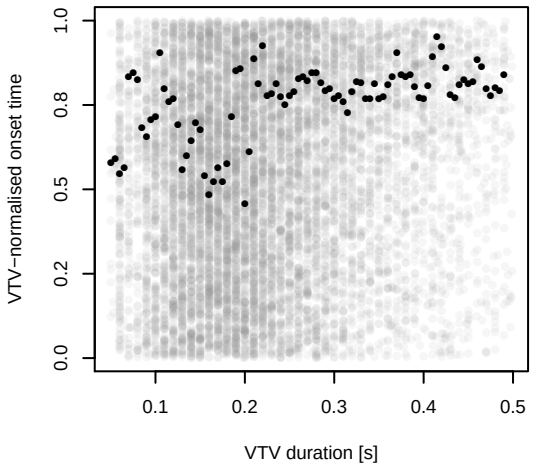

Figure 5.14: Scatterplot of VTV-normalised overlap onset time against VTV duration (grey) and mean onset time values calculated within a window of 10 ms with 5 ms step (black).

Long VTVs might thus *trigger* overlaps by virtue of being a cue to production problems or signalling an upcoming topic shift or phrase boundary. Overlaps have been indeed found to systematically occur following such events (Shriberg et al., 2001). Therefore, these overlaps are likely to be conditioned by higher levels of linguistic or interactional organisation rather than by entrainment to rhythmic patterns in interlocutor's speech. An analogous pattern is also likely to result from a tendency of overlaps to occur towards IPU ends. This issue will be discussed in greater detail in Section 5.5.1, when top-down influences of position within an IPU are considered.

By contrast, VTVs shorter than 200 ms are too brief to trigger a purely reactive behaviour. They are, in fact, shorter than the minimal vocal response time to an external stimulus, also estimated at around 200 ms (Fry, 1975). In other words, in short VTV simultaneous speech simply cannot be a reaction to a vocalic stimulus. Rather, the effect needs to be brought about by some other process. As we will see in the next section, more stable temporal patterns emerge when overlaps are preceded by sequences of regular intervals, corroborating existence of an underlying rhythmic principle.

5.4 Effects of directly preceding speech rhythm

As the median duration of overlapped VTVs in Switchboard equals 220 ms, in roughly 50% of all instances target units are too brief to warrant an explanation in terms of overlaps being produced in reaction to the preceding vowel onset. In Chapter 1.3.5 a body of evidence was reviewed which suggests a rhythmic basis for interpersonal and sensorimotor coordination. Models outlined in Chapter 3 apply these principles to coordination of speaker changes in turn-taking phonomena and claim that the exact instant of speech initiation is guided by the previous speaker's speech rhythm. The results outlined in this chapter so far are fully consistent with this hypothesis. Presently we test it more directly by investigating the link between the degree of isochrony in speech preceding an overlap onset and the emerging temporal coordination between speakers.

More precisely, as regular durational patterns should offer a more stable basis for entrainment, resulting in stronger interspeaker coupling and tighter temporal coordination, more marked deviations from a random baseline should be observed for overlaps following highly isochronous speech. This is akin to Couper-Kuhlen's (1993) claim that rhythmisation of speech by the previous speaker in the vicinity of a TRP is a precondition for a timely turn transfer. Specifically, in the present case rhythmic entrainment should be facilitated by regularly spaced vowels prior to an overlap onset.

To verify this hypothesis, normalised Pairwise Variability Index (rPVI) of three VTVs preceding an overlap onset was calculated for 3688 overlaps in Switchboard coinciding with VTVs shorter than 200 ms in overlappee's speech. rPVI (Low and Grabe, 1995) is a measure of how much neighbouring intervals differ in duration:

$$rPVI = \sum_{k=1}^{m-1} |d_k - d_{k+1}| / (m-1)$$

where d_i is the duration of the i-th interval. Low rPVI values correspond to regular interval sequences and high values are indicative of irregular patterns.

Overlap onset times were subsequently split into three equally sized classes depending on rPVI values of the preceding VTVs (i.e. on the 33rd and 66th percentiles of the rPVI values). The resulting histograms are presented in Figure 5.15. The low and mid rPVI classes are noticeably less flat than the irregular high rPVI class, and differ significantly from randomness ($p < 0.001$, 1-KS). The difference was not statistically significant for the high rPVI class ($p = 0.73$). In addition, the low rPVI class is different from the high rPVI class ($p = 0.013$, 2-KS with Bonferroni correction). The difference between the remaining categories was not statistically significant ($p = 0.28$ and $p = 0.15$ for the Low-Mid and High-Mid comparisons respectively, 2-KS with Bonferroni correction).

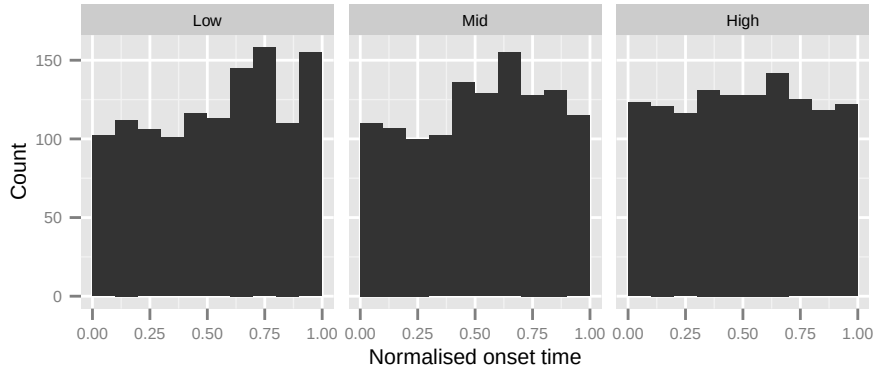

Figure 5.15: Distributions of VTV-normalised onset time for low, mid and high rPVI classes of three VTVs preceding an overlap.

The result provide support for the hypothesis that in VTVs shorter than 200 ms periodicity of directly preceding speech influences timing of overlap onsets. While rPVI is a fairly coarse measure of regularity, overlappers' behaviour differs in a consistent fashion across its values. Overlap onsets distribution is less random for more regular VTV sequences (corresponding to low rPVI

values) than for more irregular sequences (the high rPVI class). Thus, speakers can be observed to achieve more stable phase relationships between overlap onsets and regularly spaced vowel boundaries.

5.5 High-level influences

In Section 5.3 we claimed that influences of end-of-turn prediction, phrasing, hesitations, etc. on overlap timing should not be thought of as arguments in favour of the type of interspeaker adaptation investigated here. This does not mean, of course, that those phenomena do not involve coordination of interlocutors' behaviour. They obviously do. The focus of this work, however, has been on temporal orderliness at a much lower level of organisation, which is hypothesised to be brought about by rhythmic entrainment rather than mechanisms implicated in predicting ends of interlocutors' speech, reacting to their speech production problems or completing their utterances in a timely manner.

As a result, it has been assumed so far that non-random timing of overlap onsets is driven by entrainment to *sequences* of events rather than by simple reactions to specific *one-off cues*. It has been further assumed that the mechanisms involved are conditioned by properties of speech perception with little or no contribution of semantic or pragmatic factors. However, all of these are likely to interact with the low-level coordination. Indeed, some possible top-down, high-level influences have been already noted. Above all, it has been suggested that some of the observed tendencies are brought about not by fine-grained tuning of speech production and perception but by an overall increase of overlap likelihood towards turn ends. Consequently, the impact of position within an IPU is a major theme discussed in the present section. But it is by no means the only variable which can mediate interspeaker synchronisation. Below we investigate a couple of other factors, namely pragmatic influences of dialogue act type and lexical effects of word boundaries. We also study the impact of phonological syllable weight in Finnish.

5.5.1 Effects of duration and position within the IPU

Initiations of overlapping speech are not distributed uniformly across an overlappee's turn. Rather, most overlaps occur late in an IPU, a consequence of two different turn-taking strategies. On the one hand, overlaps are likely to be initiated in the vicinity of projected TRPs when the next speaker starts speaking prior to completion of the previous speaker's turn and, as a result, overlaps its final portion. On the other hand, a competitive overlap initiated

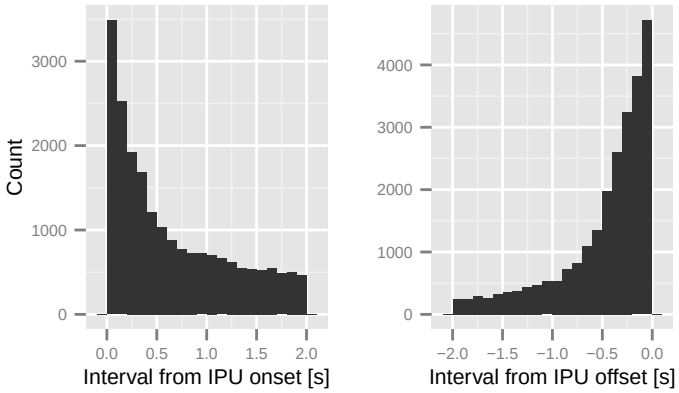

Figure 5.16: Frequencies of overlaps initiated at different intervals from IPU onsets (left) and IPU offsets (right).

in the middle of an IPU in progress might lead to its premature termination, effectively producing a similar tendency for clustering of overlaps around IPU offsets. Furthermore, a comparable but weaker increase in frequency of overlaps is also expected near IPU onsets due to both speakers breaking the silence simultaneously.

Figure 5.16 demonstrates that this is indeed the case. Shown there are frequencies of overlap initiations at different intervals from the onset and offset of an overlapped IPU. As predicted, there is a marked increase in overlap likelihood around IPU boundaries: more overlaps are produced following an IPU onset and prior to an IPU offset.

While the above is hardly surprising and follows from the mixed-initiative character of dialogue, a more fundamental question is how these tendencies interact with the fine effects reported earlier. As sequential organisation of interaction is likely to exert stronger influence on overlap timing than subtle prosodic events associated with vowel onsets, the expected pattern is a gradual skewing of the distribution of VTV-normalised overlap onset time in the vicinity of IPU boundaries[4]. Specifically, simultaneous starts should shift normalised overlap onset values to the left in VTVs early in an IPU, and terminal overlaps should shift them to the right in VTVs late in an IPU. The dependency is portrayed schematically in Figure 5.17, where VTV boundaries are overlayed on an idealised distribution of overlaps within an IPU. Any pattern specific to vowel onsets (or any other sub-IPU events, e.g. pitch accents) will be superimposed on this overall IPU-related trend. The latter might thus be thought of as providing

[4]A similar effect of IPU boundaries should be observed on IAI intervals. However, only VTV-related effects are analysed below due to a larger data set.

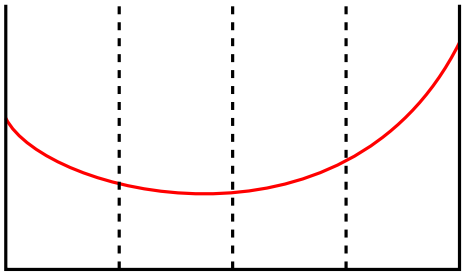

Figure 5.17: Schematic representation of the skewing effect of the distribution of overlaps within an IPU (red line) on timing of overlap onsets within VTVs (dashed lines). The effect is different depending on the position of a VTV within an IPU—stronger near IPU boundaries, weaker near IPU midpoints.

a prior distribution of overlap timing within smaller units. However, as this prior is not linear, it necessarily differs depending on the position of a VTV within an IPU. Generally speaking, stronger effects should be expected in the vicinity of IPU boundaries than near IPU midpoints, where the influence of the prior may be assumed to be negligible.

To investigate this hypothesis, overlap onsets normalised by the duration of the entire IPU are plotted against VTV-normalised onset time as a heatmap in Figure 5.18. Frequencies were scaled to 1 column-wise, the gradient thus corresponds to relative frequency of VTV-normalised overlap onset time values at different relative positions within an IPU.

As predicted, VTV-normalised onset time can be seen to increase across an IPU. In addition, the influence of IPU offsets seems to extend over a larger domain than that of IPU onsets. Somewhat worryingly, the VTV-normalised onset time rises monotonically and does not stabilise as overlaps approach IPU midpoints, which would be expected given temporal patterning independent of IPU boundary influences. However, it should be borne in mind that Figure 5.18 conflates long and short IPUs by normalising their durations, whereas the influence of turn boundaries might be more adequately described by means of absolute rather than relative position, as indicated by Figure 5.16, where VTV-onset time is plotted against the distance from an IPU onset (left) and an IPU offset (right). As the figure makes clear, the impact of IPU boundaries does not extend further than 1 second. By contrast, in short IPUs the boundary effects might be so strong as to obscure any evidence of finer organisation.

To overcome these problems, distributions of VTV-normalised onset time were analysed separately depending on whether an overlap was started within the initial second of the IPU duration (the *Early* position), the final second of the IPU duration (the *Late* position) or in-between the two regions (the *Medial*

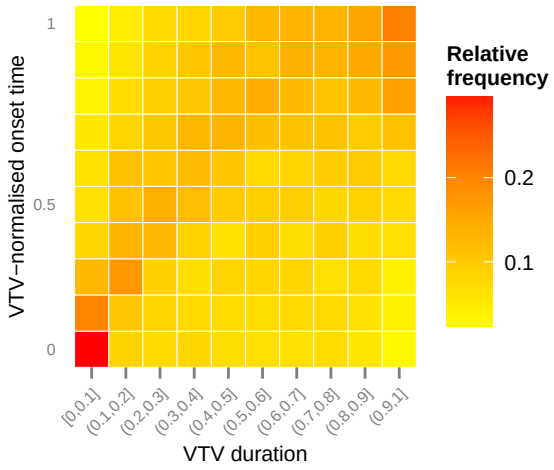

Figure 5.18: A heatmap of VTV-normalised overlap onset time against IPU-normalised overlap onset time. Frequencies were scaled to 1 column-wise.

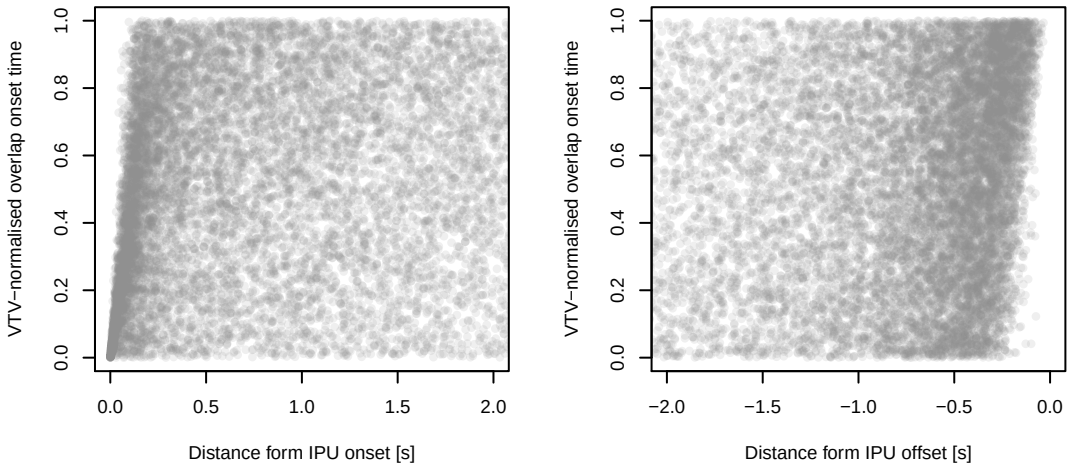

Figure 5.19: A scatterplot of VTV-normalised onset time up to two seconds following the IPU onset (left) and preceding the IPU offset (right).

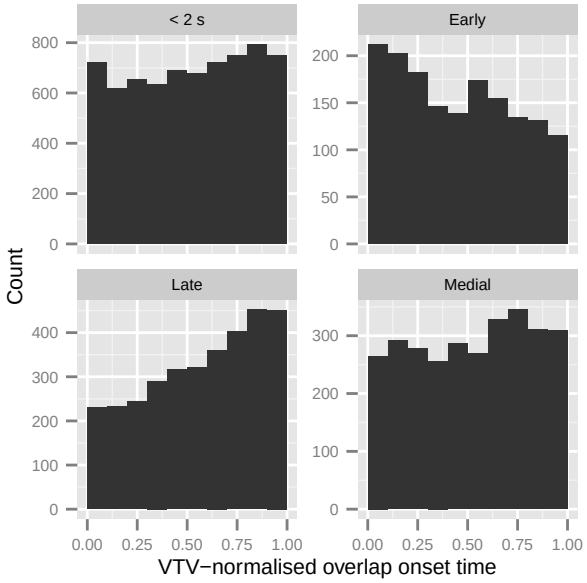

Figure 5.20: Distributions of VTV-normalised overlap onset time depending on the position of an overlap in an IPU: within the first second (*Early*), within the last second (*Late*), and separated by at least 1 second from either IPU boundary (*Medial*). Overlaps coinciding with IPUs shorter than 2 seconds are plotted separately ($< 2s$).

position). Overlaps coinciding with IPUs shorter than two seconds were assigned to a separate category. All four distributions are presented as histograms in Figure 5.20.

The *Early* and *Late* categories are consistent with the hypothesised IPU effect (see Figure 5.17). A clear tendency to pull overlap onsets in the direction of the nearest IPU boundary is manifested in the opposite skewing found in the first and last second of the IPU duration. Additionally, overlaps produced in the IPU-medial position display the familiar preference for late overlap onsets around 75% of the VTV duration. As these overlaps were initiated at least 1 second away from either IPU boundary, they are unlikely to be related to simultaneous starts and terminal overlaps. We, therefore, claim that the effect observed in the *Medial* category corresponds to interspeaker entrainment brought about by sensitivity to the detail of interlocutor's speech production rather than global conversational structure. Notably, the *Medial* category is significantly different from both a uniform distribution and the other two categories ($p < 0.001$).

IPUs shorter than 2 seconds were deemed too brief to be split into subparts in a meaningful way. It was assumed that in those cases the boundary-related effects are likely to obscure the fine temporal organisation of overlapping speech

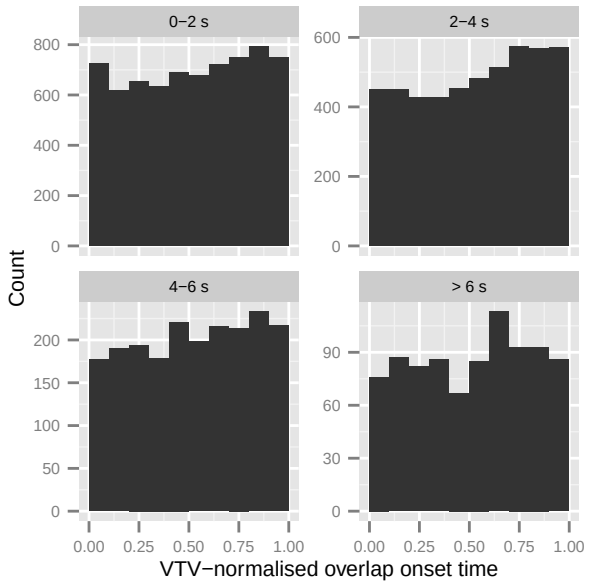

Figure 5.21: Distributions of VTV-normalised overlap onset time in increasingly long IPUs.

completely. Conversely, as IPUs get longer increasingly many overlaps should be initiated in the medial position, making IPU-medial entrainment more prominent. The relative contributions of IPU boundaries and IPU-internal rhythmic organisation to overlap timing can thus be investigated by comparing VTV-normalised overlap onset time in IPUs of increasing durations. The results, plotted in Figure 5.21, are in line with the hypothesised influence of simultaneous starts and transitional overlaps on the local organisation of overlaps around vowel onsets. The former effects are indeed more manifest in short IPUs than in long ones: as IPU durations increase, the peaks around 0 and 1 gradually disappear and give way to the IPU-internal coordination producing the late peak within a VTV. All distributions are statistically different from random ($p < 0.001$ for IPUs shorter than 6 seconds and $p = 0.042$ for IPUs longer than 6 seconds, 1-KS).

In sum, it appears that position within an IPU does to a large extent shape the temporal patterns of overlap initiation within VTVs. We claimed that the overall distribution of overlaps in an IPU could be treated as a baseline for all sub-IPU phenomena. As a consequence, to conclude presence of a statistically significant effect, the observed distributions should be more appropriately compared against a prior associated with an IPU rather than against a random uniform distribution, as has been done in the present work. The difficulty, however, lies in the fact that such a prior describes a non-linear relationship and, consequently,

changes from one VTV to the next (and from one overlap to the next). This is the case since, as shown in Figure 5.17, each VTV will span a different part of the IPU-specific distribution. The problem is further complicated by the fact that overlaps are likely to be distributed differently across IPUs of different durations. Effectively, no single prior can be established but should be constructed separately *for each data point*. Unfortunately, no statistical procedure for constructing such a prior and comparing an observed distribution against it is known to the present author. Nonetheless, non-random patterns of overlap timing in IPU-medial VTVs, where influences of IPU boundaries should be negligible, corroborates the existence of interspeaker entrainment underlying timing of overlapping speech.

5.5.2 Overlap onsets within words

The target units used thus far have all been defined by phonetically / phonologically conditioned boundaries and have abstracted from other levels of linguistic description. By contrast, in this section we investigate to what extent and in what way timing of overlaps is determined by presence of word boundaries.

The existence of lexical influences on turn-taking is supported by the observation that dialogue partners time their utterances by way of anticipating *what* the interlocutor is going to say, hinting at a role of lexico-syntactic factors behind end-of-turn prediction (de Ruiter et al., 2006). Although this claim has been put forward to explain scarcity of overlaps rather than regularities involved in their production, it is plausible that similar mechanisms might determine placement of simultaneous speech onsets. Assuming presence of lexically-mediated turn planning, overlappers might align simultaneous speech onsets with word ends. As a result, more overlaps should be initiated near word boundaries than word-medially.

With a view to validating this claim, word boundaries were extracted from the Switchboard corpus and normalised onset time was calculated for each overlap in the usual way with the exclusion of overlaps coinciding with the IPU-initial and IPU-final words. Overall, 9309 overlaps were analysed.

The distribution of word-normalised overlap onset time is plotted in Figure 5.22 alongside subject means. The distribution is statistically different from a uniform baseline ($p < 0.01$, 1-KS). At the same time, however, it is strikingly similar to that of syllable-normalised onset time (Figure 5.1), and the difference between the two is not statistically significant ($p = 0.17$, 2-KS). As it turns out, the resemblance can be explained by frequencies of overlapped words with different numbers of syllables, plotted in Figure 5.23. Clearly, almost 75% of all overlapped words are monosyllabic, with 16% of bisyllabic and 6% of trisyllabic

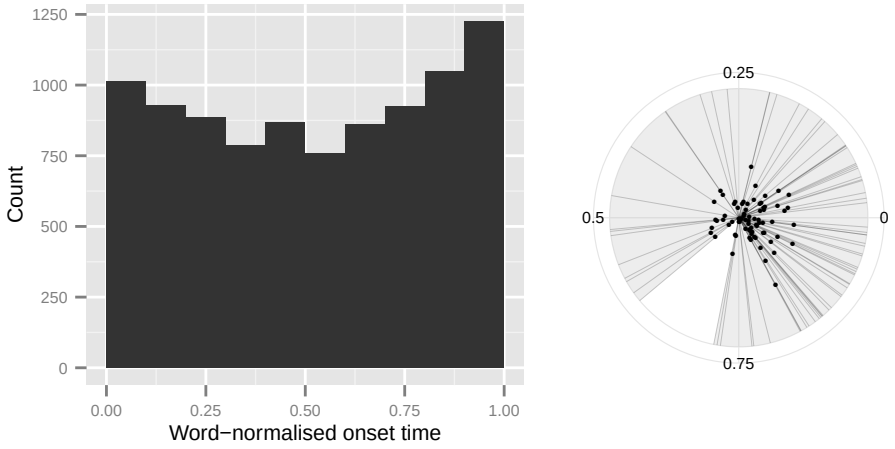

Figure 5.22: Distribution (left) and subject means (right) of word-normalised onset time in English.

words. Accordingly, Figures 5.1 and 5.22 do indeed represent to a large extent the same phenomenon. Nevertheless, because of the coincidence of the two unit types, the nature of the underlying organisation cannot be inferred.

To tease apart the interaction of syllabic and lexical boundaries, in Figure 5.24 word-normalised onset time is plotted separately for mono and polysyllabic words. Both distributions are significantly different from a random baseline ($p < 0.001$, 1-KS) and, more importantly, also different from one another ($p < 0.001$, 2-KS). While monosyllabic words not surprisingly reproduce the shape observed for syllables, polysyllabic words show a different pattern characterised by a steady rise throught their duration.

Consequently, the U-shaped pattern appears to be specific to coordination of overlap onsets around syllable boundaries and cannot be attributed to lexical influences. If the latter were true, similar distributions should be observed in mono- and polysyllabic words, which is clearly not the case. In addition, the increase in overlap frequency towards the end of polysyllabic words is more likely to reflect a non-circular dependency rather than rhythmic entrainment to lexical boundaries. For instance, it might be brought about by prediction of word ends as proposed by de Ruiter et al. (2006), providing the effect is limited to polysyllabic words only, or correspond to the distribution of overlaps within larger units, such as IPUs (see Section 5.5.1).

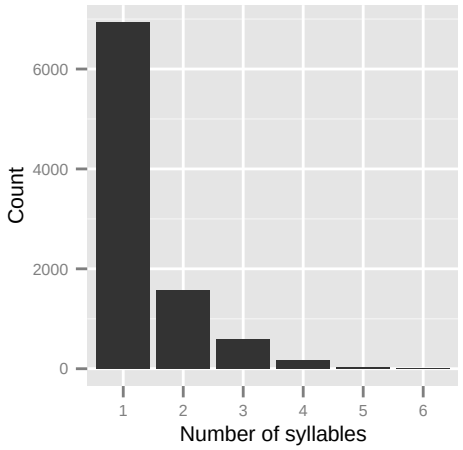

Figure 5.23: Frequencies of words with different numbers of syllables.

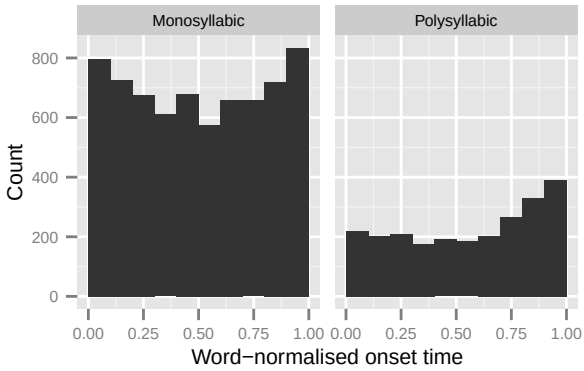

Figure 5.24: Distributions of normalised onset time within mono- and polysyllabic English words

5.5.3 Effects of dialogue act

Another plausible hypothesis regarding high-level factors influencing overlap timing pertains to pragmatic function of overlapping utterances. Namely, it could be expected that overlaps whose pragmatic function consists in expressing agreement or support of the current speaker will show evidence of better, or at any rate different, temporal alignment than other overlap kinds. We investigate this possibility by comparing rhythmic properties of different pragmatic types of overlap, operationalised by means of dialogue act categories of overlapping utterances.

Dialogue act (DA) labels of simultaneously produced utterances (both overlapper's and overlappee's) were thus retrieved for 10858 overlaps in the Switchboard corpus. As before, overlaps coinciding with IPU-initial VTVs were excluded from the analysis. VTV-normalised onset time was then calculated separately for each DA category. It was found, however, that overlaps are not distributed uniformly across DAs, resulting in very low counts in some classes. In Figure 5.25 we plot frequencies of overlaps involving different DA types of overlapper's utterances.[5] Given the effect of target unit duration demonstrated in Section 5.3, the data was additionally split on the duration of the target VTV in overlappee's speech (shorter/longer than 200 ms). As can be seen, backchannels, with over 4000 instances, make up the bulk of all cases. Several other categories (*agreement*, *opinion* and *statement*) reach moderate frequencies, with comparatively few data points in the remaining classes.

Consequently, all clearly affiliative DAs (*acknowledge*, *affirm*, *agree*, *apprec*, *backchannel*, *completion*, *downplay*, *thank*, *yes*[6]) were collapsed and compared against the remaining categories pooled together. The resulting distributions of normalised onset time within VTVs shorter than 200 ms are shown in Figure 5.26. The distributions have strikingly similar shapes and are in fact not statistically distinguishable ($p = 0.67$, 2-KS), indicating, somewhat surprisingly, that pragmatic factors (at least as far as these are reflected in DA categorisation) have little impact on termporal organisation of overlapping speech. Only the affiliative category was significantly different from a uniform distribution (1-KS, $p = 0.003$, compared to $p = 0.43$ for the remaining DAs).

Although it is difficult to formulate any specific hypothesis concerning the relationship between overlap timing and pragmatic function of *overlappees'* utterances, the same procedure was repeated for DAs of the original speaker. As can be seen in Figure 5.27 practically all instances of overlap coincide with statements or statements expressing opinion. As a result, the two categories were compared against the remaining DAs (Figure 5.28). Similar to overlappers'

[5]For an explanation of DA labels see Appendix A.

[6]See footnote 5.

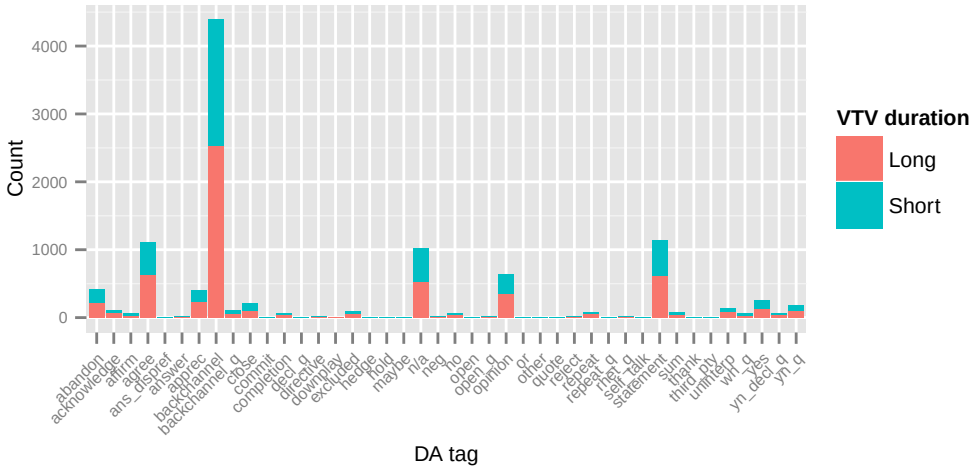

Figure 5.25: Frequencies of DA tags of overlapper's utterances, split depending on whether the duration of the target VTV in overlappee's speech exceeded 200 ms (*Long*) or not (*Short*).

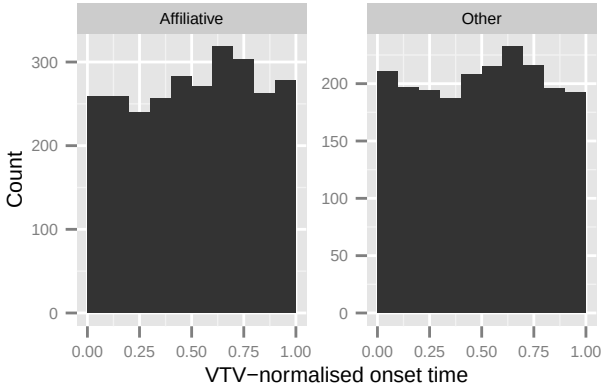

Figure 5.26: VTV-normalised overlap onset time for overlapper's affiliative and non-affiliative utterances.

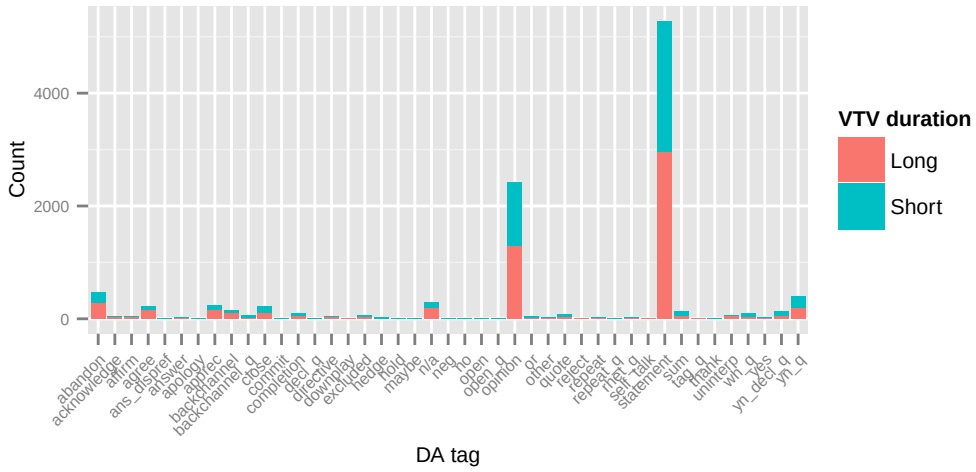

Figure 5.27: Frequencies of DA tags of overlappee's utterances, split depending on whether the duration of the target VTV in overlappee's speech exceeded 200 ms (*Long*) or not (*Short*).

utterances, however, no marked differences are discernible. Visually, the *Other* category is somewhat flatter, and statements have a secondary peak at 0 but all three exhibit a similar tendency towards late overlap onset. Neither is any of the distributions statistically distinguishable from the other two (2-KS: Statement/Opinion, $p = 0.4$; Statement/Other, $p = 0.95$; Opinion/Other $p = 0.28$). In addition, the distribution of utterances of type *Opinion* was significantly different from random (1-KS: $p = 0.008$, compared to $p = 0.068$ for statements and $p = 0.577$ for the other DA types).

In sum, we have not found any evidence for a systematic link between pragmatic function (captured by the DA categories used) and timing of overlapping utterances. This was true both of an overlapper, where clearly affiliative dialogue acts were contrasted with the remaining categories, and an overlappee. Not only were there no indications of increased synchrony in affiliative incomings, in fact no significant difference could be found whatsoever.

As the reader will recall from Chapter 3.1, Couper-Kuhlen (1993) made a related point about an association between disruptiveness and timing of overlapping speech. Specifically, she conjectured that "off-beat" or "syncopated" overlaps are heard as more turn-competitive than overlaps produced in a timely fashion. Unfortunately, a direct test of Couper-Kuhlen's (1993) hypothesis does not seem possible without an independent classification of each overlapping speech instance along the collaborative–competitive continuum. For this reason, our analysis in this section is clearly not exactly equivalent to the relationship between conversational rhythm and intrusiveness as put forward by Couper-

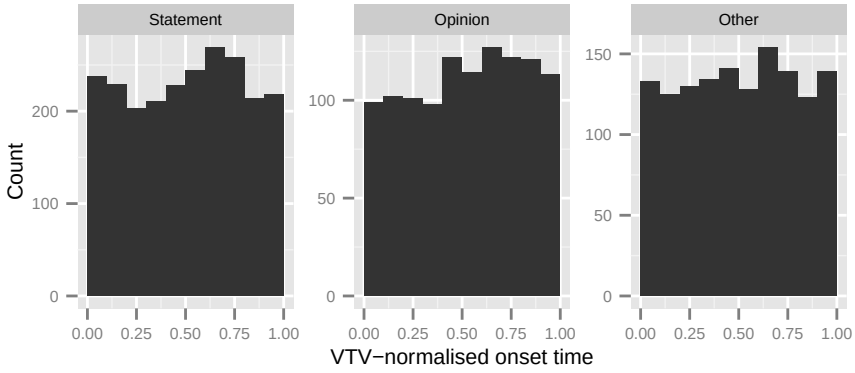

Figure 5.28: VTV-normalised overlap onset time for overlappee's statements, statements expressing opinion and other utterances.

Kuhlen (1993). In fact, her formulation seems to suggest that an utterance of *any* pragmatic type can be disruptive to varying degrees depending on its timing. In other words, rhythmic factors can outweigh the pragmatic function of an utterance. It could be even assumed that utterances with disaffiliative functions will be produced in a more rhythmic fashion precisely to compensate for and disguise their face-threatening meaning.

It is of course possible that temporal factors might override the default pragmatic meaning of an utterance, such that even the most benign and collaborative overlap be heard as intrusive, and *vice versa* (for a similar point see also Wagner et al., in press). The question begs further research. Nevertheless, in the light of the results presented in this and the previous sections, it appears that the low-level entrainment under discussion is not sensitive to lexical and/or pragmatic influences.

5.5.4 Effects of syllable weight

Even though our results point towards a universal nature of the effect in question, possible phonological influences associated with individual languages cannot be entirely ruled out. In the present case, since Finnish exhibits different phonological properties than the other three languages investigated here, it is a promising platform for comparisons.

Finnish has been described as mora-timed as opposed to syllable- or stress-timed (O'Dell et al., 2008), and contrasting *light* syllables (consisting of one mora) with *heavy* syllables (consisting of two or more morae). Assuming a language-specific and phonologically motivated effect, different timing patterns of overlap initiation could arise in each syllable type. Consequently, overlap onsets were

partitioned into two classes depending on whether the overlapped VTV was light or heavy. Even though the light/heavy distinction applies primarily to syllables, syllable-initial consonants in Finnish have no effect on syllable weight. Hence, light VTVs can be defined as consisting of a short vowel optionally followed by a single consonant (originally belonging to the following syllable), and heavy VTVs as consisting of a short vowel followed by more than one consonant or of a long vowel optionally followed by additional consonants.

Overlap onset distributions for the two VTV types are plotted in the top panel of Figure 5.29.[7] Patterns of overlap initiation are quite different in both VTV kinds. In heavy VTVs the distribution approximates a U-shape, with a slight preference for overlaps to precede vowel onsets. By contrast, most overlaps are concentrated around 75% of the light VTV duration. The tendencies are also clearly perceivable when subject means are compared across the categories (Figure 5.29, bottom panel). The mean range test was significant for heavy VTVs ($p = 0.019$) but not for light ones ($p = 0.087$). In addition, the distributions were not significantly different from one another ($p = 0.62$, 2-KS).

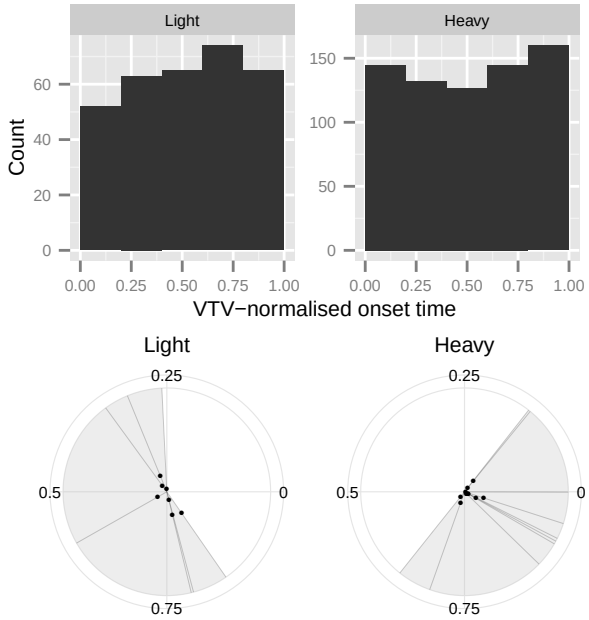

Figure 5.29: Distributions (top) and subject means (bottom) of normalised overlap onset time in light and heavy VTV intervals in Finnish.

[7] Due to small sample sizes coarser binning was used.

The observed difference hints at a role of phonologically defined syllable weight in defining coordinative dynamics. However, as indicated above, heavy and light VTVs differ substantially in segmental structure. Thus the difference in peak location between the categories in Figure 5.29 might well correspond to segmental characteristics of target units rather than to their phonologically defined weight. We explore this possibility by analysing the effect of Finnish quantity contrasts on overlap alignment.

In Figure 5.30 Finnish VTVs were further divided into four classes depending on their segmental structure: VC (a short vowel followed by a consonant), VVC (a long vowel or a diphtong followed by a consonant), VCC (a short vowel followed by a long consonant or a consonant cluster) and VVCC (a long vowel or a diphthong followed by a long consonant or a consonant cluster). Notably, even though VVC, VCC and VVCC classes all correspond to heavy VTV intervals, overlap onsets are distributed differently in each class, suggesting that phonological weight alone may not provide an adequate account of the observed patterns.

At the same time, the maximal concentration of overlaps follows roughly the expected ratio of vocalic and consonantal segment durations. In particular, in VVC intervals, in which the vocalic segment is comparatively the longest, overlaps cluster later than in VVCC intervals, in which the vowel is shorter at the expense of the consonantal portion. A similar, if weaker, VTV-medial grouping of overlaps is visible in the light VC class. Indeed, even though quantity contrasts result in more lengthening in vowels than in consonants (Lehtonen, 1970), VC and VVCC intervals are indeed expected to have comparable vocalic-to-consonantal duration ratios. Finally, in VTVs of type VCC the late peak is accompanied by another concentration early in a VTV, which is absent from the remaining distributions. The pattern is thus consistent with the expected short vocalic interval. None of the observed distributions was significantly different from a uniform one (VCC: $p = 0.872$, VVCC: $p = 0.065$, VC: $p = 0.705$, VVC: $p = 0.199$, 1-KS).

The tendency is even clearer when mean normalised onset time values for the categories are compared (Figure 5.31): overlaps occur on average later in VTVs with relatively long vowels (VVC) than in VTVs with relatively short vowels (VCC). The other types (VVCC and VC) again reach intermediate values.

The above results lend provisional support to the hypothesised link between vowel duration and the likelihood of overlap initiation. They are, however, based on very weak tendencies observed on small data samples. In addition, the vocalic-to-consonantal ratios have been inferred from Finnish quantity contrasts rather then measured from the speech signal. To directly check this hypothesis and cross-validate it on a different data set, in Figure 5.32 we plot

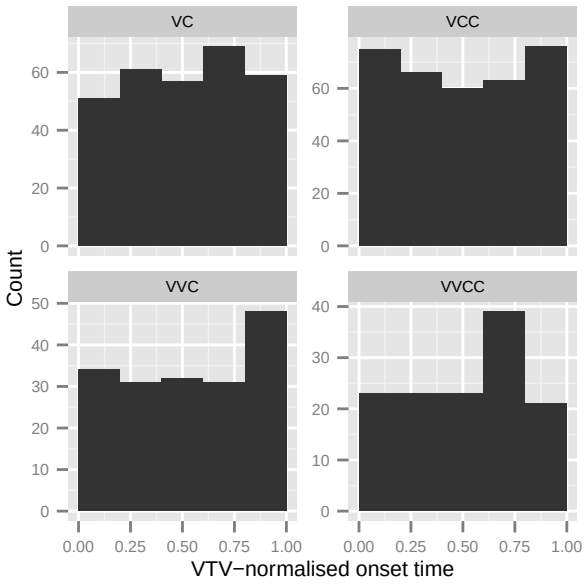

Figure 5.30: Distribution of normalised overlap onset time in VTV intervals with different segmental structure in Finnish.

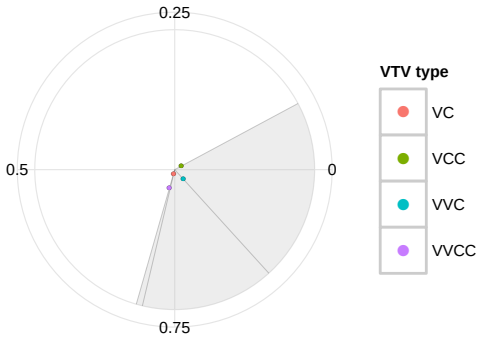

Figure 5.31: Mean normalised overlap onset time in VTV intervals with different segmental structure in Finnish. Range is represented by the shaded area.

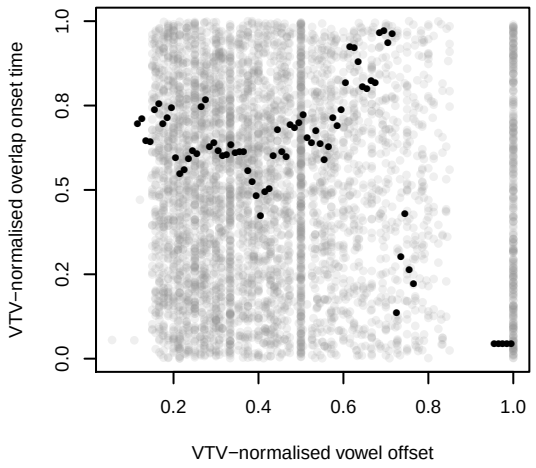

Figure 5.32: Scatterplot of normalised onset time against normalised vowel offset (grey) with mean values calculated over a series of overlapping windows (black).

VTV-normalised onset time against vowel duration, also normalised to the duration of a VTV, in the Switchboard corpus. Data points are represented as semi-transparent grey dots with mean VTV-normalised overlap onset time values calculated over a series of overlapping windows overlplotted in black. If a link between vowel duration and overlap timing were indeed present, a linear relationship between vowel offset and overlap onset time would be expected: overlaps should be initiated later in VTVs with longer vowels than in those with shorter vowels. The expectation is indeed partly borne out by the data. Average VTV-normalised overlap onset time does in fact rise when relative vowel duration increases from 40 to 70% of the total VTV duration. For shorter vowels a smaller decrease is discernible. Insufficient number of data points for vowels longer than 70% of the VTV duration precludes meaningful analysis.

To sum up, even though the evidence is somewhat weak and not entirely conclusive, vocalic segments appear to have an effect on timing of overlap initiations. Overlaps in the Finnish data were observed to change in a systematic fashion according to vocalic and consonantal quantity contrasts. Notably, the account provides an alternative to the explanation in terms of syllable weight. Specifically, a perceptually-motivated prominence-based interpretation was offered of what initially appeared to be a phonological influence. In addition, in English a direct association has been found between VTV-normalised onset

time and the percentage of VTV duration corresponding to a vocalic segment. However, the relation was only restricted to vowels covering more than half of the total VTV duration, suggesting that shorter vowels may not be perceptualy salient enough to trigger the effect.

Chapter 6

Discussion

The body of evidence presented in this work provides consistent and long-lacking empirical evidence for temporal entrainment in dialogue. Dialogue participants were demonstrated to achieve a high degree of precision in timing of overlap onsets. This tendency was observed for a number of linguistic units: Overlaps were found to coincide with syllabic boundaries in the English data but no such tendency was reproduced for French. Overlap alignment was subsequently recalculated with respect to vowel onsets, which are known to correspond to perceived syllable boundaries and to be implicated in coordination with external rhythmic stimuli. This yielded significant and comparable results for both English and French, as well as for Finnish and German: overlaps were most likely shortly before the vowel onset and least likely following it. The similarity of results for all four data sets and especially the significant outcome for French testifies to robustness of vowel onsets as opposed to phonologically-defined syllable boundaries, whose exact location is often disputable. Notably, this interpretation is also in line with the results of Beňuš (2009), who found stronger indications of entrainment for perceptually salient pitch accents than for syllables. It has been also demonstrated that overlap timing is not sensitive to pragmatic utterance function. Furthermore, given that the corpora used represent a wide spectrum of dialogue types (telephone and face-to-face, conversations between friends and strangers, with or without a predefined topic), global interaction variables seem to have little impact on the phenomenon under study.

Thus, it was hypothesised that dialogue partners become coordinated in time through entrainment to sequences of perceptual prominences in their interlocutors' speech. In other words, rhythmic organisation of perceptually salient speech events provides affordances for successful synchronisation of speaking turns.

Insofar as vocalic segments constitute one such type of prominence, the tendency to initiate overlaps before an upcoming vowel onset is consistent with this hypothesis. Further evidence was provided by Finnish VTV intervals partitioned into classes based on their segmental properties, where a tendency was observed for the location of the maximum likelihood of overlap initiation to change depending on relative durations of vocalic and consonantal portions of a VTV. Specifically, overlap onsets occur later when a vocalic segment is (relatively) the longest (as in VTVs with a long vowel followed by a single consonant) than when it is the shortest (as in VTVs with a short vowel followed by a long consonant or a consonant cluster), with the intermediate cases (VTVs with short vowels followed by a single consonant and with long vowels followed by a long consonant or a consonant cluster) falling in between the two extremes. A direct link between vowel duration and overlapping speech onset was also observed for English.

Consistent results were obtained for other kinds of prominence too. In particular, overlap onsets were found to be distributed non-uniformly within intervals between consecutive pitch accents, attesting to presence of a pitch-related effect. Specifically, the shape of the distribution indicates a decreased likelihood of overlap onsets in the vicinity of pitch accents. Since prominence in English has been described as related primarily to pitch (Fry, 1958), this finding is in line with the proposed role of perceptual prominence in guiding interspeaker entrainment.

In addition to postulating the role of perceptual prominence in temporal organisation of simultaneous speech, an attempt has been made to evaluate contributions of individual correlates of prominence to the observed effect. Subsequent analyses tried to relate the effect observed on IAIs to the previously reported VTV effects. However, the obtained results are by no means straightforward. Although accented VTVs were on the whole found to exhibit a stronger pattern than non-accented VTVs, results in Figure 5.11 suggest that the effect can be mainly attributed to duration rather than to presence of pitch accents. Generally, long VTVs exhibited a markedly non-random pattern, regardless of their accentedness. By contrast, short VTVs, whether accented or not, were not found to deviate significantly from a random baseline. Thus, there is little evidence for the impact of accentedness on timing of overlaps within VTVs. Insofar as the hypothesised link between temporal patterns in overlap onsets and prominence is correct, this is consistent with those accounts of prominence in English which have described it mainly in terms of durational features (Silipo and Greenberg, 1999)

Notably, a task related to ours was pursued by Cummins (2009b), who in a series of experiments with modified stimuli sought to assess contributions of

various acoustic features to successful synchronisation of a text read in parallel by two speakers. Although people have been observed to be extremely skilled at this task, little is known about the properties of the signal which allow such tight interspeaker coupling. However, the hypothesised importance of F_0 contour and amplitude envelope were only partly borne out, since neither feature on its own provided a sufficiently strong cue for synchronisation. Consequently, the results point towards an intricate interplay of all the factors:

> Together, these results suggest that synchronization among speakers is facilitated both by intelligibility, and by specific information within the signal, some of which may be processed in a speech-specific manner. [The stimuli used] serve to caution that the role of the amplitude envelope, that has frequently been supposed to be a principal carrier of macroscopic, rhythmic, information in the signal may be somewhat overstated. There is a complex interplay between amplitude, fundamental frequency and spectral characteristics that remains to be further clarified.

Cummins (2009b) concludes: "[c]ollectively, then, a combination of envelope modulation, F_0 and long-term spectral properties are implicated in facilitating synchronization among speakers" (Cummins, 2009b, 24). It seems likely that a similar interplay between various features might be at work in our data, and may possibly depend on the language under investigation. Nevertheless, for lack of a statistically significant result within short non-accented VTVs in the portion of the corpus labelled for pitch accents, a definite conclusion concerning the relationship between the influences of pitch accents and vocalic onsets, and the likely interaction of duration, is not possible at present. Notably (and perhaps not coincidentally), a similar challenge of separating various contributing factors has been also noted in studies of perceptual prominence (Tamburini, 2006; Wagner et al., 2012).

Moreover, even though duration seems to be the sole feature influencing timing of overlap onsets, it should be borne in mind that the number of data points in the analysed categories might have been too small to allow detecting subtler influences of vocalic segments and accentuation. Indeed, the analysis of overlap onset alignment within VTVs in the entire corpus revealed a different organisation within short and long VTVs, with a qualitative change occurring around the threshold of 200 ms. Essentially, the likelihood of overlap onset rises monotonically throughout longer intervocalic intervals, leading to a high concentration of overlaps near VTV ends. Based on manual inspection of a number of randomly selected instances, we proposed that this trend in the data might correspond to overlaps being triggered by *individual cues* in overlappee's speech, e.g. backchannel-preceding pitch contours. In addition, duration itself

might be a cue to disfluencies or an upcoming phrase boundary, both of which have been noted to frequently invite overlaps (Shriberg et al., 2001). The VTV-final increase in overlap count was also found to be associated with a general tendency to produce overlaps in the vicinity of IPU boundaries. As expected, terminal overlaps were shown to result in a strong negative skew in overlap distribution, and simultaneous starts produced a corresponding, if somewhat weaker, positive skew.

We have noted previously the puzzling lack of circularity in distributions of VTV-normalised overlap onset time manifested in discontinuities in overlap frequencies near VTV boundaries. The puzzle lies in the fact that, given a circular organisation of overlaps within VTV intervals, overlaps should be distributed symmetrically around vowel onsets. These discontinuities might be plausibly explained by the non-circular organisation of overlap onsets within IPUs or within longer VTVs corresponding to phrase-final or disfluent speech. A related trend was also observed in polysyllabic words, which might be indicative of an upcoming lexical boundary prediction.

An entirely different distribution was found in VTVs shorter than 200 ms. In these intervals overlaps were not concentrated at the very end of a VTV but clustered around 75% of the VTV duration, with similar overlap frequencies found on either side of VTV boundaries, suggesting a circular process proper. The effect was statistically significant, albeit much weaker and likely to be obscured by higher-level influences. A similar pattern was found in overlapped VTVs occurring at least one second away from IPU boundaries, where the influence of simultaneous starts and terminal overlaps should be minimal.

Importantly, the qualitative differences found imply that duration cannot fully account for the pattern observed on short VTVs. Instead, the possibility that the underlying mechanism is rhythm- rather than reaction-driven was evidenced by the influence of rhythmicity of speech preceding overlap. When overlaps were partitioned depending on regularity of preceding vocalic onsets, it became apparent that more regular sequences result in greater deviations from random overlap timing. As more regular components are by definition easier to entrain to, there appears to be a rhythmic component underlying onset placement, in line with claims made by Wilson and Wilson (2005) and Couper-Kuhlen (1993), as well as results by Buder and Eriksson (1997, 1999).

We thus concluded existence of at least two parallel processes determining timing of overlap onsets. On the one hand, overlaps are associated with specific well-defined one-off cues of phrase or turn-boundaries, feedback-inviting features, disfluencies, etc. These phenomena are more suitably described in terms of reaction and prediction, and do not belong to the domain of interspeaker entrainment as far as the latter denotes arriving at stable phase patterns be-

tween oscillatory processes. On the other hand, non-random patterns are also found at locations within overlappee's speech which cannot be easily accounted for by purely reactive or predictive processes. In those cases the underlying process appears to be rhythm-based, and satisfies the criteria of interspeaker entrainment.

More generally, accounts of temporal entrainment in terms of planning and prediction of future events do not seem to provide a satisfactory explanation for the fine temporal organisation of mid-IPU and short VTVs. Indeed, it is difficult to see how any model centred around high-level and centrally-governed planning could deal both with the time scale and with the variability of the phenomena discussed here (see also Lenneberg 1971), especially at lower levels of prosodic hierarchy, such as syllables and VTVs. For this reason rhythm-based accounts of entrainment and the notion of coordinative structure might prove especially instructive. A dynamical hypothesis, along the lines discussed in Chapter 1.3.5 would hold that just as muscles form functional complexes specialised in performing a specific activity (e.g. grasping or throwing), dialogue partners enter into close-knit relationships and form a single functionally-defined speaking unit. As the reader will remember, the body of work reviewed in Sections 1.2 and 1.3, especially the results of Shockley et al. (2003), Schmidt et al. (1990), Marsh et al. (2009) and Cummins (2009b) indicate that coordinative structures are by no means necessarily tied to a single organism but can subsume distinct individuals and be mediated by convergent speech patterns.

A similar dependency might underlie overlap onsets timing with previous speaker's speech providing (and enforcing) a common temporal frame, shared by both participants, who are bound to operate within its limits. As in other instances of coordination, the emerging coupling between dialogue partners necessarily leads to reduction in complexity of the coordinated unit, thereby imposing constraints on the temporal patterns of speech initiation. In the present case the dimensional compression introduces preferences (attractors) for some phase dependencies between sequences of speech events and simultaneous speech onsets resulting in higher concentration of overlaps at certain locations. This, in turn, is in line with the enactive perspective on cognition; as Shockley et al. (2009, p. 315) point out: "characterising interpersonal coordination as a cross-person coordinative structure may require a different conceptualization of cognition. Cognition, from a dynamical systems perspective, may be more usefully understood as a set of constraints on action."

Indications of another well-known property of coordinative structures also exist. Namely, given that in Chapter 5.4 non-random patterns were observed for both low and mid rPVI classes, some compensatory mechanisms must be at work. Perfect regularity does not seem to be a prerequisite for successful en-

trainment as speakers are able to deal with moderate deviations from isochrony. At the same time, once the departures become too large (as in the most irregular high rPVI category), coordination deteriorates. This suggests presence of a threshold above which the yoking between participants is loosened.

As pointed out by Cummins (2009a), the concept of affordance implies a strong coupling between perception and action. It reflects a basic symmetric correspondence between what is perceived and a possible course of action. In the present case it translates into a link between the perceived speech of the interlocutor and one's own simultaneous articulatory activity. Such a link is in itself nothing new in phonetic sciences. Similar claims have been made repeatedly since the 1950s, starting with Liberman's motor theory of speech perception (Liberman, 1957; Liberman and Mattingly, 1985) stating that "the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands." More recently, findings in neurosciencies, such as those of mirror neurons (Rizzolatti and Craighero, 2004), which are activated both when performing and observing an action, or of increased potentials in listeners' articulatory muscles while listening to speech produced predominantly with the same articulators (Fadiga et al., 2002) have provided additional support for this hypothesis. However, an important difference between these findings and the present one exists lies in the fact that the link implied in our results requires no assumptions about either representational levels (as in Liberman and Mattingly's (1985) theory) nor with neural or muscle activation (as in Rizzolatti and Craighero's (2004) and Fadiga et al.'s (2002) work). What is truly extraordinary about the results reported in the previous chapter is that the coupling can be observed directly in the outward behaviour itself. Speech *production* of one speaker is quite literally tied to speech production of his or her interlocutor. To the best of this author's knowledge no such dependency has so far been demonstrated in the domain of spontaneous, non-controlled dialogue. Additionally, the present findings are in agreement with the recent claim that the motor cortex, which has long been postulated as the locus of interaction between speech perception and production, might be more likely responsible for recording interlocutor's speech rhythm, and tied to sensorimotor coordination involved in production of speaking turns (Scott et al., 2009).

In demonstrating a rhythmic organisation of turn-taking our results are of course in agreement with the tenets of the models discussed in Chapter 3. At the time we mentioned that, attractive as they are, empirical evidence supporting rhythm- and entrainment-based accounts of turn-taking is scarce. We also noted that such state of affairs might be at least partly caused by the methodological problems of studying rhythm of speaker changes accompanied

by silence, which necessitates resorting to measures based on extrapolation of some interval derived from the previous speaker's turn. However, as O'Dell et al., 2012 point out, it is not clear what the characteristics of the extrapolated sequences should be: whether they should correspond to means calculated over the previous speaking turn, to the duration of the final unit of interest within that turn (e.g. the last syllable) or whether it should reproduce its durational variability. Indeed, the fact that earlier studies (see Chapter 3) have failed to find evidence of interspeaker entrainment for "clean" speaker changes suggests that dialogue partners do not arrive at some shared dialogue rhythm by means of straightforward extrapolation of speech event sequences. This is hardly surprising given that in speech stream the left context is updated with every new unit produced, resulting in deviations from any pattern established so far. In the light of the results presented above it appears that whatever the mechanism governing overlap production, it cannot consist solely in repetition of some averaged unit. Instead, it appears to be sensitive to and capable of compensating for moment-by-moment changes in speech regularity. That information is obviously unavailable during stretches of silence. In that respect, it is instructive that Beňuš (2009) observed more synchronous patterns across overlapped turn transition than across other speaker change configurations. Furthermore, contrary to the initial expectations, with the exception of English syllables no clear in- or anti-phase patterns have been found. Phase values of about 0.75 and 0.6 have been observed for intervals between consecutive vowel onsets and pitch accents respectively, pointing towards complex dependencies between levels of linguistic and prosodic organisation, the exact nature of which is not fully understood at present. At any rate, a coupled oscillator model, especially a hierarchically organised one akin to that proposed by O'Dell et al. (2012) seems much more suitable for modelling the phenomena under investigation than a simple linear or constant extrapolation strategy.

Chapter 7

Conclusions and future work

This work has attempted to bring together three distinct strands of research: speech convergence, turn-taking and synchronisation among speakers. We have argued that overlapping speech is an excellent testing ground for such efforts as it allows tracing of temporal relationships between events in interlocutors speech directly in the data. Indeed, using a simple measure quantifying timing of overlap onset with respect to arbitrary events in the interlocutor's speech (e.g. vowel onsets, syllabic boundaries, pitch accents, etc.), we have observed non-random patterns of overlap initiation at various levels of prosodic hierarchy, a finding which has so far eluded a rigorous empirical analysis. At the same time, overlap timing has been demonstrated to be insensitive to other features, such as dialogue act categories of overlapping utterances. Consequently, a hypothesis was put forward according to which overlaps are produced with reference to perceptually salient events in speech. Particular attention has been paid to timing of overlap onsets with respect to consecutive vowel onsets but compatible results have been obtained for pitch accents. The relationship between the two was investigated but was inconclusive due to insufficient sample sizes. In addition, prominence-related tendencies have been found to co-exist with higher levels of interactional organisation, most importantly associated with IPU boundaries. The resulting pattern is, therefore, a combination of local and global factors.

The results are also highly instructive about cognitive mechanisms underlying interpersonal coordination. While average durations of intervals between pitch accents are sufficiently long to consider a reactive or predictive basis for overlap production, VTV intervals are simply too short and varied in duration to grant such a possibility. Instead, in the light of sensitivity of overlap timing to regularity of directly preceding speech, we have argued that the effect is more adequately described by speakers getting entrained to sequences of perceptually salient events in their interlocutors' speech. Drawing on the framework of coordination dynamics, we have further proposed that the resulting

organisation can be adequately modelled as a coordinative structure, a functionally defined emergent unit coupled by the underlying task and imposing behavioural constraints on its component parts. Therefore, the observed pattern should not be conceived of in terms of one individual trying to predict or react to actions of his or her interlocutor. Rather, it is brought about by the constraining nature of speech rhythm, leading to emergence of preferences for certain temporal relationships between individuals' speech (or, more precisely, articulatory movements). Consequently, it can be regarded as a very tangible example of the long-postulated perception/production link: perception of speech prior to overlap quite literally restrains the observed moment of producing an overlap. Notably, similar to other examples of coordinative structures, the mechanism has been demonstrated to be able to compensate for moderate deviations from perfect periodicity (isochrony), hence accounting for variability in speech production.

Naturally, a number of issues have remained unsolved. Above all, as pointed out in Chapter 5.5.1, given the strong influence of IPU boundaries on overlap initiation patterns, a statistical method for inferring non-random distribution of overlaps should preferably compare it against the IPU-associated prior rather than against a flat uniform distribution, as has been done here. However, such a procedure would be far from straightforward. As VTVs at different points within an IPU have different prior distributions of overlap probability due to increasing skewness when moving from the IPU midpoint towards the boundaries, a different prior needs to be associated with each analysed data point. Unfortunately, such a contingency has not been foreseen by standard statistical methods.

More generally, the combination of IPU-internal entrainment with the global IPU-specific overlap patterns calls for nonstandard statistical treatment. As we pointed out repeatedly, our data is in principle circular: whatever the unit under consideration, the ends of the scale denote essentially the same point (vowel onset, pitch accent, etc.), and the observed distribution is repeated in contiguous units. Nonetheless, the discontinuities around unit boundaries testify to presence of a parallel, non-circular process, which has been interpreted as tied to the domain of an IPU. While methods for analysing circular data exist (Batschelet, 1981), they do not commonly account for an amalgam of circular and non-circular data. Moreover, many of the more advanced methods known from non-directional statistics, such as regression models with random terms, do not exist for circular measures. Out of necessity, this work has resorted to a mixture of circular and non-circular procedures, and to analysing one factor at time (e.g. duration, accentedness, effect of dialogue act categories, etc.). This is obviously suboptimal as it does not allow tracking of interactions and

interdependencies between studied quantities. However, in view of the specific character of the analysed data, no technique seemed entirely appropriate. It is hoped that some alternative, e.g. in the form of Bayesian statistics, could alleviate the situation and facilitate drawing a more comprehensive picture of mechanisms shaping the temporal organisation of overlapping speech.

In addition to different analytic methods, experimental and modelling efforts could help separate different factors contributing to interspeaker entrainment. Production experiment with overlapping speech providing a high degree of control are not trivial. Overlap is, after all, a phenomenon characteristic of unrestrained dialogue and is difficult to elicit in laboratory conditions, for example using pre-recorded sounds samples with varying rhythmic or durational patterns. Similarly, instructing participants to interrupt filtered or otherwise heavily modified speech to pinpoint features necessary for synchronisation is at odds with boundary conditions of dialogue. A solution might be offered by human-computer interaction studies, in which system output is modified to match specific hypotheses. It should be borne in mind, however, that human participants might be less inclined to synchronise with a machine than with a fellow human, as suggested by Kawasaki et al.'s (2013) results. For this reason a real-time modification of interlocutors' speech along the lines of Morris (1971) might be more desirable but limitations of fine-grained signal alternation performed online pose an obvious problem. Otherwise, following Szczepek-Reed (2010), entrainment patterns between native speakers of languages with different rhythmic properties could be studied. "Interactions" between speakers unfamiliar with one another's languages might prove particularly instructive in highlighting the language specific and universal aspects of entrainment. Another alternative could be sought in designs following the *parasocial consensus sampling* paradigm (Huang et al., 2010), in which subjects are asked to interact with one side of a pre-recorded interaction. However, given a non-interactive nature of this task, subjects may be reluctant to initiate simultaneous speech altogether.

An issue entirely untouched in the present work is that of dialogue participants' sensitivity to the investigated phenomena. It is not clear whether different overlap patterns are at all perceivable by humans and whether they correlate with other aspects of interaction, such as speaker's likeability, task performance or rapport between dialogue partners. These questions should be addressed by appropriate perceptual tests.

Dynamical modelling is another possible approach to hypothesis testing. As already mentioned, a hierarchical coupled oscillator model accounting for interdependencies between various prosodic levels could be used. Alternatively, an account grounded in strictly local effects of perceptual prominence could

be advanced by modelling such stretches of speech as attractors and repellers, pulling overlaps away or towards them with varying forces.

Naturally, the modelling and experimental pursuits do not mean that the possibilities of corpus analysis have been already exhausted. Several lines of enquiry could still be followed. For instance, coordination in dialogue is known to fluctuate, with periods of alignment and misalignment (De Looze et al., 2011). Thus, it could be expected that adaptation on one level might be facilitated and reinforced by adaptation on other levels. Consequently, more consistent patterns of overlap initiation should be observed when interlocutors adapt on other features, for example F_0 or intensity. Studying dependencies between overlap timing and temporal entrainment along other dimensions, such as body sway, gesturing or breathing, could be especially fruitful. Relatedly, an implicit assumption of this work has been that whatever the mechanism determining overlap initiation might be, they remain constant over the entire dialogue. Even though preliminary analyses seemed to corroborate this claim, it need not necessarily be the case. A more dynamic, time-dependent method, possibly combined with tracking other interaction states, could help shed more light on this problem.

On a more basic level, this work has been restricted to timing of simultaneous *speech onsets*. Thus, the possibility sketched out by Couper-Kuhlen (1993), namely that certain landmarks are aligned *throughout stretches of overlapping speech* has not been thoroughly explored. Initial analytic attempts were not particularly encouraging but the issue naturally deserves more attention. Lastly, influence of factors such as speakers' familiarity with each other, availability of visual cues, etc. could be also studied in greater detail.

In sum, we believe that the implications of the present work constitute another convincing argument in favour of granting overlapping speech its legitimate place in accounts of human interaction. It is hoped that investigation of temporal organisation of overlaps can be a fruitful enterprise not only for studies of dialogue but can also inform models of coordination between action and perception and, by extension, advance the understanding of human cognition.

Appendices

Appendix A

NXT-Switchboard dialogue act inventory

NXT DA tag	Category
abandon	Adandoned or Turn-Exit
acknowledge	Response Acknowledgment
affirm	Affirmative non-yes answers
agree	Agree/Accept
ans_dispref	Dispreferred answers
answer	Other answers
apology	Apology
apprec	Appreciation
backchannel	Backchannel
backchannel_q	Backchannel as question
close	Conventional-closing
commit	Offers, Options & Commits
completion	Collaborative Completion
decl_q	Declarative Wh-Question
directive	Action-directive
downplay	Downplayer
excluded	Excluded - bad segmentation
hedge	Hedge
hold	Hold before response
maybe	Maybe/Accept-part
neg	Negative non-no answers
no	No answers
open	Conventional-opening

NXT DA tag	Category
open_q	Open-Question
opinion	Statement-opinion
or	Or-Clause
other	Other
quote	Quotation
reject	Reject
repeat	Repeat-phrase
repeat_q	Signal-non-understanding
rhet_q	Rhetorical-Questions
self_talk	Self-Talk
statement	Statement-non-opinion
sum	Summarize/Reformulate
tag_q	Tag-Question
thank	Thanking
third_pty	3rd-party-talk
uninterp	Uninterpretable
wh_q	Wh-Question
yes	Yes answers
yn_decl_q	Declarative Yes-No-Question
yn_q	Yes-No-Question

Bibliography

Allen, G. D. (1972). The location of rhythmic stress beats in English: An experimental study I & II. *Language and Speech 15*, 72–100.

Allen, J. and M. Core (1997). Draft of Damsl: Dialogue Act Markup in Several Layers. `ftp://ftp.cs.rochester.edu/pub/packages/dialog-annotation/manual.ps.gz`.

Altmann, U. (2010). Investigation of movement synchrony using windowed cross-lagged regression. In A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt (Eds.), *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues*, Volume 6800 of *Lecture Notes in Computer Science*, pp. 335–345. Berlin: Springer.

Andersen, P. A. (1992). Excessive intimacy: An account analysis of behaviors, cognitive schemata, affect, and relational outcomes. In *Proceedings of the International Conference on Personal Relationships*, Orono, ME.

Argyle, M. and M. Cook (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Argyle, M. and J. Dean (1965). Eye-contact, distance and affiliation. *Sociometry 28*(3), 289–304.

Barbosa, P. A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In *Speech Prosody 2002*, pp. 163–166.

Bargh, J. A. and T. L. Chartrand (1999). The unbearable automaticity of being. *American Psychologist 54*(7), 462–479.

Batschelet, E. (1981). *Circular statistics in biology*. New York: Academic Press.

Bavelas, J. B., A. Black, C. R. Lemery, and J. Mullett (1986). "I show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology 50*(2), 322–329.

Bavelas, J. B., L. Coates, and T. Johnson (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication 52*(3), 566–580.

Beattie, G. W. (1982). Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica 39*(1-2), 93–114.

Bell, L., J. Boye, J. Gustafson, and M. Wirén (2000). Modality convergence in a multimodal dialogue system. In *Proceedings of Götalog 2000, Fourth Workshop on the Semantics and Pragmatics of Dialogue*, Göteborg, pp. 29–34.

Bell, L., J. Gustafson, and M. Heldner (2003). Prosodic adaptation in human-computer interaction. In *Proceedings of ICPhS*, pp. 2453–2456.

Beňuš, Š. (2009). Are we 'in sync': turn-taking in collaborative dialogues. In *Proceedings of Interspeech 2009*, Brighton, U.K., pp. 2167–2170.

Beňuš, Š. (2011). Adaptation in turn-initiations. In A. Esposito (Ed.), *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, Lecture Notes in Computer Science [LNAI], pp. 72–80. Berlin / Heidelberg: Springer-Verlag.

Beňuš, Š., A. Gravano, and J. Hirschberg (2011). Pragmatic aspects of temporal accomodation in turn-taking. *Journal of Pragmatics 41*(12), 3001–3027.

Bernieri, F. J., J. Reznick, and R. Rosenthal (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology; Journal of Personality and Social Psychology 54*(2), 243–253.

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. New York: Pergamon Press.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy (2008). Le CID — Corpus of Interactional Data — Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues 49*(3), 1–30.

Bertrand, R. and R. Espesser (2000). About speech overlaps. Prosodic cues contribution in predicting a change of speaker. In *Proceedings of Prosody 2000*, Kraków, Poland.

Branigan, H. P., M. J. Pickering, and A. A. Cleland (2000). Syntactic co-ordination in dialogue. *Cognition 75*(2), B13–B25.

Branigan, H. P., M. J. Pickering, J. Pearson, J. F. McLean, and C. I. Nass (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 186–191.

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, Philadelphia, PA, pp. 41–44.

Brennan, S. E. and H. H. Clark (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology 22*(6), 1482–1493.

Brown, P. and S. C. Levinson (1987). *Politeness*. Cambridge: Cambridge University Press.

Buder, E. H. (1991). A nonlinear dynamic model of social interaction. *Communication Research 18*(2), 174–198.

Buder, E. H. and A. Eriksson (1997). Prosodic cycles and interpersonal synchrony in american english and swedish. *Proceedings of Eurospeech 1997 1*, 235–238.

Buder, E. H. and A. Eriksson (1999). Time-series analysis of conversational prosody for the identification of rhythmic units. In *Proceedings of the 14th International Congress of Phonetic Sciences*, Volume 2, San Francisco, CA, pp. 1071–1074.

Buder, E. H., A. S. Warlaumont, D. K. Oller, and L. B. Chorna (2010). Dynamic indicators of mother-infant prosodic and illocutionary coordination. In *Proceedings of the 5th International Conference on Speech Prosody*.

Bull, M. (1996). An analysis of between-speaker intervals. In *Proceedings of the Edinburgh Linguistics Conference '96*, Edinburgh, pp. 18–27.

Burgoon, J. K., L. A. Stern, and L. Dillman (2007). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge: Cambridge University Press.

Byrne, D. (1971). *The Attraction Paradigm*. New York: Academic Press.

Campbell, N. (2007). Approaches to convesational speech rhythm: speech activity in two-person telephone dialoges. In *Proceedings of ICPhS XVI*, Saarbrucken, pp. 343–348.

Cappella, J. N. and J. Green (1982). A discrepancy-arousal explanation of mutual influence in expressive bahavior for adult and infant-adult interaction. *Communication Monographs 49*, 89–114.

Cappella, J. N. and J. Green (1984). The effects of distance and individual differences in arousability on nonverbal involvement: A test of discrepancy-arousal theory. *Journal of Nonverbal Behavior 8*(4), 259–286.

Cappella, J. N. and S. Planalp (1981). Talk and silence sequences in informal conversations iii: Interspeaker influence. *Human Communication Research 7*(2), 117–132.

Çetin, Ö. and E. Shriberg (2006). Overlap in meetings: ASR effects and analysis by dialog factors, speakers, and collection site. In S. Renals, S. Bengio, and J. Fiskus (Eds.), *Machine Learning for Multimodal Interaction (Third Internation Workshop, MLMI 2006, Bethesda, MD)*, Volume 4299 of *Lecture Notes in Computer Science*, Berlin, pp. 212–224. Springer.

Chapple, E. D. (1939). Quantitative analysis of the interaction of individuals. *Proceedings of the National Academy of Sciences of the United States of America 25*(2), 58–67.

Chapple, E. D. (1970). *Culture and biological man: Explorations in behavioral anthropology*. New York: Holt, Rinehart and Winston.

Chartrand, T. L. and J. A. Bargh (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology 76*(6), 893–910.

Condon, W. and W. Ogston (1971). Speech and body motion synchrony of the speaker-hearer. In D. L. Horton and J. J. Jenkins (Eds.), *The Perception of Language*, pp. 150–184. Colombus, OH: Charles E. Merrill.

Cooper, S. (2011). Frequency and loudness in overlapping turn onset by welsh speakers. In *Proceedings of ICPhS XVII*, Hong Kong, pp. 516–519.

Coulston, R., S. Oviatt, and C. Darves (2002). Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Volume 4, pp. 2689–2692.

Couper-Kuhlen, E. (1991). A rhythm-based metric for turn-taking. In *Proceedings of the 12th International Congress of Phonetic Sciences*, Aix-en-Provence, France, pp. 275–278.

Couper-Kuhlen, E. (1993). *English speech rhythm: form and function in everyday verbal interactions.* Amsterdam: John Benjamins.

Cummins, F. (2009a). Rhythm as an affordance for the entrainment of movement. *Phonetica 66*(1–2), 15–28.

Cummins, F. (2009b). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics 37*(1), 16–28.

Cummins, F. (2011a). Coordination, not control, is central to movement. In A. Esposito, A. M. Esposito, R. Martone, V. C. Müller, and G. Scarpetta (Eds.), *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, Volume 6456 of *Lecture Notes in Computer Science*, pp. 243–255. Springer.

Cummins, F. (2011b). Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience 5*(170), 1–9.

Cummins, F. (2012). Oscillators and syllables: a cautionary note. *Frontiers in Psychology 3*, 1–2.

De Looze, C., C. Oertel, S. Rauzy, and N. Campbell (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. In *Proceedings of ICPhS XVII*, Hong Kong, pp. 1294–1297.

de Ruiter, J., H. Mitterer, and N. J. Enfield (2006). Predicting the end of a speaker's turn; a cognitive cornerstone of conversation. *Language 82*(3), 515–535.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review 3*(4), 357.

Dressler, R. A., E. H. Buder, and M. P. Cannito (2009). Rhythmic patterns during conversational repairs in speakers with aphasia. *Aphasiology 23*(6), 731–748.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology 23*(2), 283–292.

Duncan, S. and D. W. Fiske (1977). *Face-to-face interaction: Research, methods, and theory*. Hillsdale, NJ: Erlbaum.

Edlund, J. and M. Heldner (2005). Exploring prosody in interaction control. *Phonetica 62*(2–4), 215–226.

Edlund, J., M. Heldner, S. Al Moubayed, A. Gravano, and J. Hirschberg (2010). Very short utterances in conversation. In *Proceedings of Fonetik 2010*, Lund, Sweden, pp. 11–16.

Edlund, J., M. Heldner, and J. Hirschberg (2009). Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*, Brighton, UK, pp. 2779–2782.

Edlund, J. and A. Hjalmarsson (2012). Is it really worth it? Cost-based selection of system responses to speech-in-overlap. In *Proceedings of the IVA 2012 workshop on Realtime Conversational Virtual Agents (RCVA 2012)*, Santa Crux, CA.

Fadiga, L., L. Craighero, G. Buccino, and G. Rizzolatti (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience 15*(2), 399–402.

Fant, G., A. Kruckenberg, J. Liljencrants, and S. Hertegård (2000). Acoustic-phonetic studies of prominence in swedish. *Speech Transmission Laboratory–Quarterly Status and Progress Report 41*(2-3), 1–52.

Ferrer, L., E. Shriberg, and A. Stolcke (2002). Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Boulder, CO.

Finlayson, I., M. Corley, and R. Lickley (2011). Alignment in rate of speech: Evidence from a corpus of dialogue. In *Proceedings of AMLaP 2011*, Paris, pp. 67.

French, P. and J. Local (1983). Turn-competitive incomings. *Journal of Pragmatics 7*, 17–38.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and speech 1*(2), 126–152.

Fry, D. B. (1975). Simple reaction-times to speech and non-speech stimulti. *Cortex 11*(4), 355–360.

Fusaroli, R., B. Bahrami, K. Olsen, A. Roepstorff, G. Rees, C. Frith, and K. Tylén (2012). Coming to terms quantifying the benefits of linguistic coordination. *Psychological science 23*(8), 931–939.

Fusaroli, R., J. Rączaszek-Leonardi, and K. Tylén (2013). Dialog as interpersonal synergy. *New Ideas in Psychology xxx*, xxx.

Garrod, S. and A. Anderson (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition 27*, 181–217.

Giles, H., J. Coupland, and N. Coupland (1991). Accomodation theory: Communication, context and consequence. In H. Giles, J. Coupland, and N. Coupland (Eds.), *Contexts of Accomodation. Developments in applied sociolinguistics*, Chapter 1, pp. 1–69. Cambridge: Cambridge University Press.

Godfrey, J. J., E. Holliman, and J. McDaniel (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, CA, pp. 517–520.

Goldman, J.-P., M. Avanzi, A. Lacheret-Dujour, A.-C. Simon, et al. (2007). A methodology for the automatic detection of perceived prominent syllables in spoken French. In *Proceedings of Interspeech 2007*, Antwerp, Belgium, pp. 91–120.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American sociological review 25*(2), 161–178.

Gravano, A. and J. Hirschberg (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language 25*(3), 601–634.

Gravano, A. and J. Hirschberg (2012). A corpus-based study of interruptions in spoken dialogue. In *Proceedings of Interspeech 2012*.

Gries, S. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research 34*(4), 365–399.

Haken, H., J. A. S. Kelso, and H. Bunz (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics 51*(5), 347–356.

Hayes, D. P. and L. Cobb (1979). Ultradian biorhythms in social interaction. In *Of speech and time: Temporal speech rhythms in interpersonal contexts*, pp. 57–70. Hillsdale, NJ: Erlbaum.

Heins, R., M. Franzke, M. Durian, and A. Bayya (1997). Turn-taking as a design principle for barge-in in spoken language systems. *International Journal of Speech Technology 2*, 155–164.

Heldner, M. (2011). Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. *Journal of Acoustical Society of America 130*(1), 508–513.

Heldner, M. and J. Edlund (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics 38*(4), 555–568.

Heldner, M., J. Edlund, and J. Hirschberg (2010). Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech 2010*, Makuhari, Japan, pp. 3054–3057.

Heldner, M., J. Edlund, A. Hjalmarsson, and K. Laskowski (2011). Very short utterances and timing in turn-taking. In *Proceedings of Interspeech 2011*, pp. 2837–2840.

Howes, C., P. G. T. Healey, and M. Purver (2010). Tracking lexical and syntactic alignment in conversation. In *Proceedings of CogSci 2010*, Portland, OR, pp. 2004–2009.

Huang, L., L.-P. Morency, and J. Gratch (2010). Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, Volume 1, Toronto, Canada, pp. 1265–1272.

Inden, B., Z. Malisz, P. Wagner, and I. Wachsmuth (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. In N. Miyake, D. Peebles, and R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society Austin, TX: Cognitive Science Society*, Austin, TX, pp. 1721–1726. Cognitive Science Society.

Inden, B., Z. Malisz, P. Wagner, and I. Wachsmuth (2013). Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In J. Epps, F. Chen, S. Oviatt, K. Mase, A. Sears, K. Jokinen, and B. Schuller (Eds.), *Proceedings of the 15th International Conference on Multimodal Interaction (ICMI 2013)*, Sydney, Australia.

IPDS (2006). The Kiel Corpus of Spontaneous Speech, vol 4, Video Task Scenario: Lindenstrasse. DVD#1.

Jaffe, J. and S. Feldstein (1970). *Rhythms of dialogue*. New York: Academic Press.

Jefferson, G. (1973). A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica 9*(1), 47–96.

Jefferson, G. (1983). Another failed hypothesis: Pitch/loudness as relevant to overlap resolution. Technical report, Tilburg University, Department of Language and Literature.

Jefferson, G. (1984). Notes on some orderlinesses of overlap onset. In V. D'Urso and P. Leonardi (Eds.), *Discourse analysis and natural rhetoric*, pp. 11–38. Cleup Editore.

Jessen, M., K. Marasek, K. Schneider, and K. Claßen (1995). Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German. In *Proceedings of ICPhS XIII*, Volume 4, pp. 428–431.

Kawasaki, M., Y. Yamada, Y. Ushiku, E. Miyauchi, and Y. Yamaguchi (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports 3*, 1–8.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT press.

Kelso, J. A. S., K. G. Holt, P. N. Kugler, and M. T. Turvey (1980). On the concept of coordinative structures as dissipative structures: II. Empirical lines of convergence. In R. J. Stelmach, George E. (Ed.), *Tutorials in Motor Behavior*, pp. 49–70. Amsterdam: North-Holland.

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta psychologica 32*, 101–125.

Kleinke, C. L. (1986). Gaze and eye contact: a research review. *Psychological bulletin 100*(1), 78–100.

Komatani, K., T. Kawahara, and H. G. Okuno (2007). Analyzing temporal transition of real user's behaviors in a spoken dialogue system. In *Proceedings of Interspeech 2007*, Antwerp, Belgium, pp. 183–186.

Komatani, K., T. Kawahara, and H. G. Okuno (2008). Predicting ASR errors by exploiting barge-in rate of individual users for spoken dialogue systems. In *Proceedings of Interspeech 2008*, Brisbane, Australia, pp. 183–186.

Kousidis, S. (2010). *A Study of Accomodation of Prosodic and Temporal Features in Spoken Dialogues in View of Speech Technology Applications*. Ph. D. thesis, Dublin Institute of Technology, Dublin.

Kousidis, S. and D. Dorran (2009). Monitoring convergence of temporal features in spontaneous dialogue speech. In *Proceedings of the 1st Young Researchers Workshop on Speech Technology University*, Dublin, Ireland.

Kousidis, S., D. Dorran, C. McDonnell, and E. Coyle (2009). Time series analysis of acoustic feature convergence in human dialogues. In *Proceedings of SPECOM 2009*, St. Petersburg, Russia, pp. 91–96.

Kurtić, E., G. J. Brown, and B. Wells (2013). Resources for turn competition in overlapping talk. *Speech Communication 55*, 721–743.

Laskowski, K., M. Heldner, and J. Edlund (2012). On the dynamics of overlap in multi-party conversation. In *Proceedings of Interspeech 2012*, Portland, OR.

Lee, C.-C., M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and N. Shrikanth (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In *Proceedings of Interspeech 2010*, Kahuhari, Japan, pp. 793–796.

Lee, C.-C., S. Lee, and S. S. Narayanan (2008). An analysis of multimodal cues of interruption in dyadic spoken interactions. In *Proceedings of Interpspeech 2008*, Brisbane, Australia, pp. 1678–1681.

Lee, C.-C. and S. Narayanan (2010). Predicting interruptions in dyadic spoken interactions. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 5250–5253.

Lehtonen, J. (1970). *Aspects of quantity in standard Finnish*. Jyväskylä: Jyväskylän yliopisto.

Lenneberg, E. H. (1971). The importance of temporal factors in behavior. In D. L. Horton and J. J. Jenkins (Eds.), *The Perception of Language*, pp. 174–184. Colombus, OH: Charles E. Merrill.

Lennes, M. (2009). Segmental features in spontaneous and read-aloud finnish. In V. de Silva and R. Ullakonoja (Eds.), *Phonetics of Russian and Finnish. General Introduction. Spontaneous and Read-aloud speech*, pp. 145–166. Frankfurt am Main: Peter Lang.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Levitan, R., A. Gravano, and J. Hirschberg (2011). Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland, OR, pp. 113–117.

Levitan, R. and J. Hirschberg (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011*, Florence, Italy, pp. 3081–3084.

Lewandowski, N. (2012). *Talent in nonnative phonetic convergence*. Ph. D. thesis, Universität Stuttgart, Stuttgart.

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America 29*, 117–123.

Liberman, A. M. and I. G. Mattingly (1985). The motor theory of speech perception revised. *Cognition 21*(1), 1–36.

Lindblom, B. and P. MacNeilage (2011). Coarticulation: A universal phonetic phenomenon with roots in deep time. In *Proceedings of Fonetik 2011*, Stockholm, Sweden, pp. 41–44.

Low, E. and E. Grabe (1995). Prosodic patterns in Singapore English. In *Proceedgins of the XIIth ICPhS*, Volume 3, Stockholm, Sweden, pp. 636–639.

Makri-Tsilipakou, M. (1994). Interruption revisited: Affiliative vs. disaffiliative intervention. *Journal of Pragmatics 21*(4), 401–426.

Malisz, Z. and P. Wagner (2011/2012). Acoustic-phonetic realisation of Polish syllable prominence: A corpus study. In D. Gibbon, D. Hirst, and N. Campbell (Eds.), *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem*, Volume 14/15 of *Speech and Language Technology*, pp. 105–114. Poznań: Polish Phonetic Association.

Markus-Kaplan, M. and K. J. Kaplan (1984). A bidimensional view of distancing: Reciprocity versus compensation, intimacy versus social control. *Journal of Nonverbal Behavior 8*(4), 315–326.

Marsh, K. L., M. J. Richardson, and R. C. Schmidt (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science 1*(2), 320–339.

Matarazzo, J. D., M. Weitman, G. Saslow, and A. N. Wiens (1963). Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior 1*(6), 451–458.

McAuley, J. D. (1995). *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing*. Ph. D. thesis, Indiana University, Bloomington, IN.

McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language and Hearing Research 44*(1), 128–143.

Meltzer, L., W. N. Morris, and D. P. Hayes (1971). Interruption outcomes and vocal amplitude: Explorations in social psychophysics. *Journal of Personality and Social Psychology 18*(3), 392–402.

Monada, L. and F. Oloff (2011). Gesture in overlap: The situated establishment of speakership. In G. Stam and M. Ishino (Eds.), *Integrating Gestures: The interdisciplinary nature of gesture*, pp. 321–337. Amsterdam: John Benjamins.

Moore, R. K. (2012). Finding rhythm in speech: A response to Cummins. *Empirical Musicology Review 7*(1-2), 36–44.

Morris, W. N. (1971). Manipulated amplitude and interruption outcomes. *Journal of Personality and Social Psychology 20*(3), 319–331.

Morton, J., S. M. Martin, and C. Frankish (1976). Perceptual centers (p-centers). *Psychological Review 83*, 405–408.

Murata, K. (1994). Intrusive or co-operative? A cross-cultural study of interruption. *Journal of Pragmatics 21*(4), 385–400.

Murray, S. O. and L. H. Covelli (1988). Women and men speaking at the same time. *Journal of Pragmatics 12*(1), 103–111.

Natale, M. (1975a). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology 32*(5), 790–804.

Natale, M. (1975b). Social desirability as related to convergence of temporal speech patterns. *Perceptual and Motor Skills 40*(3), 827–830.

Nenkova, A., A. Gravano, and J. Hirschberg (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, Stroudsburg, PA, pp. 169–172.

Niederhoffer, K. G. and J. W. Pennebaker (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology 21*(4), 337–360.

Norwine, A. C. and O. J. Murphy (1938). Characteristic time intervals in telephonic conversation. *Bell System Technical Journal 17*(2), 281–291.

O'Connell, D. C., S. Kowal, and E. Kaltenbacher (1990). Turn-taking: A critical analysis of the research tradition. *Journal of psycholinguistic research 19*(6), 345–373.

O'Dell, M., M. Lennes, and T. Nieminen (2008). Hierarchical levels of rhythm in conversational speech. In *Proceedings of Speech Prosody 2008*, Campinas, Brazil, pp. 355–358.

O'Dell, M., M. Lennes, and T. Nieminen (2012). Modeling turn-taking rhythms with oscillators. *Linguistica Uralica 3*, 218–227.

O'Dell, M., M. Lennes, S. Werner, and T. Nieminen (2007). Looking for rhythms in conversational speech. In *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pp. 1201–1204.

O'Dell, M. and T. Nieminen (1999). Coupled oscillator model of speech rhythm. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 1075–1078.

O'Dell, M., T. Nieminen, and L. Mietta (2012). Integrating turn-taking behavior in a hierarchically organized coupled oscillator model. In *Book of Abstracts of Perspectives on Rhythm and Timing*, Glasgow, U.K., pp. 46.

Oertel, C., M. Włodarczak, J. Edlund, P. Wagner, and J. Gustafson (2012). Gaze patterns in turn-taking. In *Interspeech 2012*, Portland, OR.

Oreström, B. (1983). *Turn-taking in English conversation*. Krieger Pub Co.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America 119*(4), 2382–2393.

Pardo, J. S. (2012). Reflections on phonetic convergence: Speech perception does not mirror speech production. *Language and Linguistics Compass 6*(12), 753–767.

Patterson, M. L. (1976). An arousal model of interpersonal intimacy. *Psychological Review 83*(3), 235.

Patterson, M. L. (1983). *Nonverbal behavior: A functional perspective*. New York: Springer.

Pickering, M. J. and S. Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioural and Brain Sciences 27*(2), 169–190.

Portele, T., B. Heuft, C. Widera, P. Wagner, and M. Wolters (2001). Perceptual prominence. In *Speech and Signals. Aspects of Speech Synthesis and Automatic Speech Recognition. Festschrift dedicated to Wolfgang Hess on his 60th birthday*, Volume 69 of *Forum Phoneticum*, pp. 97–116. Frankfurt a.M.: Hektor.

Rapp-Holmgren, K. (1971). A study of syllable timing. *Speech Transmission Laboratory–Quarterly Status and Progress Report 1*, 14–19.

Raux, A. and M. Eskenazi (2009). A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, CO, pp. 629–637.

Reitter, D., F. Keller, and J. D. Moore (2006). Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, New York, pp. 1–4.

Richardson, D. C., R. Dale, and N. Z. Kirkham (2007). The art of conversation is coordination common ground and the coupling of eye movements during dialogue. *Psychological science 18*(5), 407–413.

Richardson, D. C., R. Dale, and J. M. Tomlinson (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science 33*(8), 1468–1482.

Richardson, D. C., R. Dale, J. M. Tomlinson, and H. H. Clark (2008). What eye believe that you can see: Conversation, gaze coordination and visual common ground. In J. Ginzburg, P. Healey, and Y. Sato (Eds.), *Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue (LONDIAL'08)*, London, pp. 83–90.

Richardson, M. J., K. L. Marsh, R. W. Isenhower, J. R. L. Goodman, and R. C. Schmidt (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human movement science 26*(6), 867–891.

Rizzolatti, G. and L. Craighero (2004). The mirror-neuron system. *Annu. Rev. Neurosci. 27*, 169–192.

Rączaszek-Leonardi, J. (2009). Symbols as constraints. the structuring role of dynamics and self-organization in natural language. *Pragmatics & Cognition 17*(3), 653–676.

Rączaszek-Leonardi, J. and J. A. S. Kelso (2008). Reconciling symbolic and dynamic aspects of language: Toward a dynamic psycholinguistics. *New Ideas in Psychology 26*(2), 193–207.

Roloff, M. E. (1987). Communication and reciprocity within intimate relationships. In R. L. Street, Jr. and J. N. Cappella (Eds.), *Interpersonal processes: New directions in communication research*, pp. 11–38. Beverly Hills, CA: Sage.

Sacks, H., E. A. Schegloff, and G. Jefferson (1974). A simplest systematics for the organization of turn-taking for conversation. *Langauge 50*(4), 696–735.

Schegloff, E. A. (2000). Overlapping talk and the organisation of turn-taking for conversation. *Language in Society 29*(1), 1–63.

Schegloff, E. A. (2002). Accounts of conduct in interaction: Interruption, overlap and turn-taking. In J. H. Turner (Ed.), *Handbook of Sociological Theory*, pp. 287–321. New York: Kluwer Academic/Plenum Publishers.

Schmidt, R. C., C. Carello, and M. T. Turvey (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance 16*(2), 227.

Schmidt, R. C., M. J. Richardson, C. Arsenault, and B. Galantucci (2007). Visual tracking and entrainment to an environmental rhythm. *Journal of Experimental Psychology: Human Perception and Performance 33*(4), 860–870.

Scott, S. K., C. McGettigan, and F. Eisner (2009). A little more conversation, a little less action – candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience 10*(4), 295–302.

Selfridge, E. O. and P. A. Heeman (2009). A bidding approach to turn-taking. In *Proceedings of the International Workshop on Spoken Dialogue Systems*.

Selfridge, E. O. and P. A. Heeman (2010). Importance-driven turn-bidding for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 177–185.

Shockley, K., A. A. Baker, M. J. Richardson, and C. A. Fowler (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology: Human Perception and Performance 33*(1), 201–208.

Shockley, K., D. C. Richardson, and R. Dale (2009). Conversation and coordinative structures. *Topics in Cognitive Science 1*(2), 305–319.

Shockley, K., M.-V. Santana, and C. A. Fowler (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance 29*(2), 326–332.

Shriberg, E., A. Stolcke, and D. Baron (2001). Observations on overlap: Findngs and implications for automatic processing of multi-party conversation. In *Proceedings of EUROSPEECH*, pp. 1359–1362.

Silipo, R. and S. Greenberg (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, Volume 3, pp. 2351–2314.

Steininger, S., F. Schiel, and K. Louka (2001). Gestures during overlapping speech in multimodal human-machine dialogues. In *Proceedings of International Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy.

Stivers, T., N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. Yoonf, and S. C. Levinson (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America 106*(26), 10587–10592.

Stoyanchev, S. and A. Stent (2009). Lexical and syntactic priming and their impact in deployed spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, Boulder, Colorado, pp. 189–192.

Street, Jr., R. L. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research 11*(2), 139–169.

Ström, N. and S. Seneff (2000). Intelligent barge-in in conversational systems. In *Proceedings of Interspeech 2000*, Beijing, China, pp. 652–655.

Szczepek-Reed, B. (2010). Speech rhythm across turn transitions in cross-cultural talk-in-interaction. *Journal of Pragmatics 42*(4), 1037–1059.

Tamburini, F. (2006). Reliable prominence identification in English spontaneous speech. In *Proceedings of Speech Prosody 2006*, Dresden, Germany.

Tamburini, F. and C. Caini (2005). An automatic system for detecting prosodic prominence in american english continuous speech. *International Journal of Speech Technology 8*, 33–44.

Tannen, D. (1994). Gender and discourse. In *Interpreting Interruption in Conversation*, Chapter Interpreting Interruption in Conversation, pp. 53–82. Oxford University Press.

Tannen, D. (2005). *Conversational Style: Analyzing Talk among Friends*. Oxford: Oxford University Press.

ten Bosch, L., N. Oostdijk, and L. Boves (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication 47*(1–2), 80–86.

ten Bosch, L., N. Oostdijk, and J. P. de Ruiter (2004). Turn-taking in social talk dialogues: temporal, formal and functional aspects. In *Proceedings of SPECOM 2004*, St. Petersburg, pp. 454–461.

Wagner, P., Z. Malisz, B. Inden, and I. Wachsmuth (2013). Interaction phonology – A temporal co-ordination component enabling representational alignment within a model of communication. In I. Wachsmuth, J. de Ruiter, P. Jaecks, and S. Kopp (Eds.), *Alignment in Communication: Towards a New Theory of Communication*, Advances in Interaction Studies, pp. 109–132. Amsterdam: Benjamins.

Wagner, P., F. Tamburini, and A. Windmann (2012). Objective, subjective and linguistic roads to perceptual prominence. how are they compared and why? In *Proceedings of Interspeech 2012*, Portland, OR.

Ward, A. and D. Litman (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*, Farmington, PA.

Ward, N. and S. Nakagawa (2004). Automatic user-adaptive speaking rate selection. *International Journal of Speech Technology 7*(4), 259–268.

Ward, N. G. and S. K. Mamidipally (2008). Factors affecting speaking-rate adaptation in task-oriented dialogs. In *Proceedings of the 4th International Conference on Speech Prosody*.

Warner, R. M. (1979). Periodic rhythms in conversational speech. *Language and Speech 22*(4), 381–396.

Warner, R. M., D. Malloy, K. Schneider, R. Knoth, and B. Wilder (1987). Rhythmic organization of social interaction and observer ratings of positive affect and involvement. *Journal of Nonverbal Behavior 11*(2), 57–74.

Welkowitz, J., S. Feldstein, M. Finklestein, and L. Aylesworth (1972). Changes in vocal intensity as a function of interspeaker influence. *Perceptual and Motor Skills 35*(3), 715–718.

Wells, B. and S. Macfarlane (1998). Prosody as an interactional resource: Turn-projection and overlap. *Language and Speech 41*(3/4), 265–294.

Wilson, M. and T. P. Wilson (2005). An oscillator model of the timing of turn taking. *Psychonomic Bulletin and Review 12*(6), 957–968.

Wilson, T. P., J. M. Wiemann, and D. H. Zimmerman (1984). Models of turn taking in conversational interaction. *Journal of Language and Social Psychology 3*(3), 159–183.

Wilson, T. P. and D. H. Zimmerman (1986). The structure of silence between turns in two-party conversation. *Discourse Processes 9*(4), 375–390.

Włodarczak, M. and P. Wagner (2013). Effects of talk-spurt silence boundary thresholdson distribution of gaps and overlaps. In *Proceedings of Interspeech 2013*, Lyon, France, pp. 1434–1437.

Yang, F. and P. A. Heeman (2010). Initiative conflicts in task-oriented dialogue. *Computer Speech and Language 24*(2), 175–189.

Yang, L.-c. (1996). Interruptions and intonation. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP 96)*, Volume 3, pp. 1872–1875.

Yngve, V. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, Chicago, pp. 567–577.

Yuan, J., M. Liberman, and C. Cieri (2007). Towards an integrated understanding of speech overlaps in conversation. In *Proceedings of ICPhS XVI*, Saarbrücken, Germany, pp. 1337–1340.

Zimmerman, D. H. and C. West (1975). Sex roles, interruptions and silences in conversation. *Language and sex: Difference and dominance 105*, 105–129.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man–Machine Studies 34*(4), 527–547.