A COMPUTATIONAL MODEL FOR UNSUPERVISED CHILDLIKE SPEECH ACQUISITION

by
Holger Brandl

Submitted to the
Technische Fakultät
Universität Bielefeld

in partial fulfillment of the requirements for the degree of a

Doktor Ingenieur

thesis supervisor	thesis supervisor
Dr.-Ing. Frank Joublin	Dr.-Ing. Britta Wrede
Honda Research Europe GmbH	Universität Bielefeld

July 2009

Acknowledgments

This thesis would not have been possible without Doreen, my beloved girlfriend and proud mother of my little sunshine Simon. I would like to thank her and my parents for supporting me throughout all the years of my studies and all their interest in my research.

I thank my supervisor Frank Joublin for his constant support, his always honest feedback, more inspiring ideas than I could ever implement, and for pushing my work forward by pluckily questioning many things I've always taken for granted. I would like to express my deep thanks to my supervisor Britta Wrede for supporting me with all her knowledge about speech recognition and infant development, and her always positive, warm, and charming attitude whenever we met to shape my fuzzy ideas into a thesis.

I thank my PhD-fellows Miguel Vaz, Xavier Domont, Bjoern Schölling, Irene Ayllon and Claudius Gläeser for all the endless but always highly inspiring discussions, countless *Blockrunden*, their sometimes sharp but always friendly and constructive feedback, and all the fun we had together. Big thanks go to Jens Schmüdderich for lifting me over the *Audiotellerrand* into the wonderful world of multi-modal pattern recognition, the bracing hours in the gym, and his brave attempts to make our life as PhD-students even more comfortable.

I would like to thank Tobias Rodemann for refocusing me whenever I drifted too far away from my topic, and also for comforting me in hours of BBCM-despair. I would say many thanks to Martin Heckmann for sharing his deep understanding of speech processing systems with me whenever I was confused, for his wonderful ability to condense fuzzy discussions to the essence of matter, and for supporting my favorite time series modeling technique as long as we've not found the perfect and ultimate replacement solution.

Many thanks go to the ALIS-team for forcing my model from a beautiful annotated corpus-universe into the rough seas of human-robot interaction. Especially I would like to thank Christian Goerick who has raised my view on system design to a new and thrilling level. I would like to thank Gerhard Sagerer, Edgar Körner, Andreas Richter and Franz Kummert for making this thesis possible - surely somewhere in between Offenbach and Bielefeld, but always feeling at home and welcome in both locations.

Finally, I would like to express my thank all reviewers of this thesis for their endless fights against my not yet converged English syntax model, and their mindfulness to reveal all the typos I've really worked hard to hide. It were their questions that helped me to capture at least a short shimmering glance on what a model for infant-inspired speech acquisition could look like.

Dresden, July 2009
Holger Brandl

Summary

Speech understanding requires the ability to parse spoken utterances into words. But this ability is not innate and needs to be developed by infants within the first years of their life. So far almost all computational speech processing systems neglected this bootstrapping process. Here we propose a model for early infant speech structure acquisition implemented as layered architecture comprising phones, syllables and words. Our model processes raw acoustic speech as input and aims to capture its structure unsupervised on different levels of granularity.

Most previous models assumed different kinds of innate language-specific predispositions. We drop such unlikely assumptions and propose a model that is developmentally plausible. We condense findings from developmental psychology down to a few basic principles that our model is aiming to reflect on a functional level. By doing so our proposed model learns the structure of speech by a multitude of coupled self-regulated bootstrapping processes.

We evaluate our model on speech corpora that have some of the properties of infant-directed speech. To further validate our approach we outline how the proposed model integrates into an embodied multi-modal learning and interaction framework running on Honda's Asimo robot. Finally, we propose an integrated model for speech structure and imitation learning through interaction, that enables our robot to learn to speak with an own voice.

Contents

List of Figures

CHAPTER 1

Introduction

The holy grail of speech recognition research is to build a system that automatically acquires the structure and meaning of spoken language. However, state of the art automatic speech recognition (ASR) frameworks are designed to detect predefined words using predefined grammar models. Even after decades of intensive research those systems still fall behind humans in terms of recognition performance, reliability, adaptivity and robustness. Over the last decade speech recognition research focused on tuning existing frameworks, without major advances in its core concepts for speech unit modeling, decoding and parameter estimation.

A promising way to overcome structural problems of automatic speech recognition, is to learn from the way humans perceive language. Human speech abilities have fascinated scientists and philosophers since long ago, but are still far from being understood. Hence, researchers attempted to understand the developmental processes that infants live through while unraveling the code of language. By doing so, the hope is to reveal what makes humans start from nothing but basic sound reception abilities and finally master the tremendously complex structure of language. Especially computational models have helped to come closer to this goal by providing means to validate or to reject hypotheses about the underlying principles and processing schemes.

Since digitalization of scientific research many computational models for speech acquisition have been proposed. However, as detailed out in this thesis existing models provide no full explanation for the process of speech acquisition.

First, most models tackle the problem of speech acquisition in the symbolic domain only, although it is not clear how and whether such approaches can be generalized to the acoustic domain. The symbol sequences used as input for such models are commonly enriched with stress and phonotactic information, even though such cues are can be hardly assumed to be innately available to the young learners.

Second, many models rely on some kind of innate representation, which is often at the level of syllables. But because syllables depend strongly on the language to be learned, it is not clear how such approaches could be extended to become valid models for language acquisition as observed in infants.

Third, existing models abstract away the complexity of language down to very small vocabularies. To some extent this is necessary and valid to simplify evaluation, but often it is due to the used computational toolkits that do not scale up accordingly.

Finally and most importantly, many models for speech acquisition tend to make unlikely assumptions about innate speech processing abilities of infants [Chr98, p. 14]. This manifests itself in a different ways as we will detail out later. For instance, some models make severe assumptions about language-specific properties of the tutoring language, which includes stress cues, assumptions about the syllabic structure or syntax models.

The motivation behind this thesis is to face these challenges by designing a developmentally plausible approach for acoustic speech structure acquisition. We present a novel framework that attempts to focus on representations and processing schemes which are coherent with findings from developmental psychology. Its basic idea is to reveal the structure of spoken language in terms of a (partially) hierarchical speech representation.

The proposed framework processes acoustic speech without any annotation or cue-enrichment as input. It is organized to incrementally capture structural constraints of the speech signal. To cope with inherently noisy speech patterns we use Hidden Markov Models and statistical transition models common in ASR systems. By doing so, our model is able to profit from the tremendous body of work and knowledge in this field.

In contrast to most existing works, we exclusively implement developmental principles that psychologists and linguists believe to affect infant speech development. This focus on *developmental plausibility* renders our attempt to speech structure acquisition into a novel computational model of the infant's speech acquisition process. Following Occam's razor we first evaluated the most straightforward word acquisition approach, which is to use utterances of word length to bootstrap new words and to apply the principle of subtraction to learn also words which do not appear as isolated utterances [Bra08]. This approach failed, because length is not a reliable cue for word segmentation. Hence, we investigated principles like metric segmentation strategies, transitional probabilistic models, or the stress constraints, that are believed to play a role in the early infant speech structure acquisition (cf. chp. 2). All these mechanisms depend on a decomposition of the speech input into syllable segments. But because syllables strongly depend on the language they cannot be assumed to be innate. Although there are promising results for syllable acquisition and segmentation on symbolic corpora, it remains a challenging task to bootstrap a syllable representation from raw acoustic speech. This is because only few syllables appear as isolated utterances in spoken language and it is not clear how infants distinct these from other short poly-syllabic phrases. However, as discussed in chapter 4, there is a broad agreement that phonotactics, which are the rules that restrict how phones are assembled to form syllables, play the central role for syllable learning.

Our model implements a *bottom-up* approach for speech structure acquisition. Making no assumptions about the input language, we first propose how to capture the phonotactics of an arbitrary tutoring language. A phonotactic parser then attempts to reveal the syllabicity of speech segments. Mono-syllabic segments define the input to the syllable learning process. The emerging syllable representation allows to decompose the speech signal into syllabic units. A residual learning scheme allows to gain further training samples from incompletely syllabified utterances to speed up the learning process. Our model links syllables and words by integrating two other important aspects of infant speech development: Residual learning using subtraction, and statistical learning that exploits co-occurrence patterns in the input language. Based on these cues we formulate a novel and powerful lexical learning mechanism, that is able to reveal mono- and also poly-syllabic words from arbitrary input languages.

A major contribution of this work is to glue different developmental principles together to form a unified computational model for acoustic speech structure acquisition. This we realize by a *three-layered architecture* which was designed to capture the structure of speech on different granularity scales. Based on findings from developmentally psychology these layers are organized in terms of *phones*, *syllables* and *words*. Each layer involves a set of relatively simple data-driven processing modules. Thereby the idea is to combine unsupervised and supervised speech segmentation methods to bootstrap a model-based speech representation. Essential to this approach are regulatory feedback loops that modulate the different learning processes. First, our framework acquires a phone-representation including a phonotactic model. Second, based on the syllabic constraints implied by these learned phonotactics and input speech obeying some properties of infant-directed speech, a syllable representation becomes bootstrapped. Finally, our framework acquires a word lexicon by implementing the above mentioned lexical learning mechanism. Technically, our system can be defined as a cascade of HMM-based speech unit spotting instances which rely on incomplete speech unit representations.

Categorization is essential to perception. But how can a category model emerge from just a sequence of stimuli, especially as speech units do not appear as released patterns over time, but are mostly embedded into an utterance context? Furthermore, adjacent patterns may influence each other. Thus, because of the complex and noisy structure of speech, snapshot learning techniques are not suited to model speech acquisition. Instead it is necessary to develop methods that derive successively more complex representations of the speech structure. We propose a self-regulated *incremental syllable clustering process* that relies on the *principle of subtraction* to extract syllabic training segments. Technically we implement this idea by applying the set of models already learned at a time instance to analyze the speech input. If a subset of these models is found to match a part of the speech stream with high confidence, the corresponding segmentation can be used to generate further training segments.

A major difference between symbolic and acoustic speech processing is the role of segmentation. In the symbolic domain the set of possible segmentation points is restricted by the granularity of the used speech symbols. But for acoustic speech this number is unlimited which converts the problem of finding segment boundaries into the problem of matching a given speech representation against the speech signal. Thus, speech classification and speech segmentation become the same process. None precedes the other. However, the role of segmentation has often been neglected in the literature. In this thesis, we show that segmentation accuracy is mandatory to bootstrap a word representation for an arbitrary input language.

Beside innate mechanisms, parental support plays an important role in many cultures for speech development. Most prominent - as detailed out in chapter 2 – is the use of child-directed speech for tutoring. This special speaking mode, emphasizes structural patterns in the tutoring language, by modulating speaking rate, pitch, syntax and other cues. Hence, to evaluate our system, it is reasonable to assume a similarly structured input for bootstrapping.

We evaluate our model by simulating the infant's learning process. We process large amounts of speech and observe how speech representation emerges. To benchmark the latter we propose to use synthetic speech generators which create statistically constrained utterances with the necessary acoustic variability. This allows us to assess the quality of the bootstrapped speech structure models with common benchmarks from machine learning and ASR development.

Speech is embedded into a complex framework comprising other communication modes, context, and semantics. So the question is, whether it is reasonable to investigate a solely speech signal driven approach for speech structure acquisition. As there is no ultimate answer to this question yet, we follow a twofold approach. First, we assume that speech structure can be learned data-driven solely from the acoustic input signal. As the design of our architecture does not rely on additional non-speech cues, we consider this to be a natural first step. Furthermore, it simplifies the evaluation of the model, as multi-modal corpora are much more harder to be evaluated than uni-modal ones. Second, we show that our model embeds seamlessly into an embodied multi-modal framework for semantic learning.

Related to embodiment are possible links between speech production and speech perception. We follow the same argument here, by designing and evaluating our framework without any links to speech production. But to prove its validity, we describe subsequently how our model embeds into a developmentally plausible framework for interaction-driven speech perception and production learning. The latter has been a joint project with Miguel Vaz [Vaz09a].

1.1 Outline

The first step towards a developmentally plausible model for speech structure acquisition is to extract and discuss computational constraints and requirements for such a system. This includes the possible representations, processing schemes as well as the computational toolkit to glue the different elements together. Thus, the remainder of this work is organized as follows. In chapter 2 we review findings from developmental psychology about the speech acquisition process as observed in infants. There we aim to highlight and condense the most relevant principles of speech structure acquisition. Subsequently in chapter 3 we summarize the pattern recognition methods that our system has evolved from. To ease the understanding of related speech acquisition models we have split our literature review into two separate chapters. First, we review approaches dealing with symbolic speech acquisition in chapter 4. Second, chapter 5 synopsizes methods that acquire acoustic speech units on different levels of granularity.

Chapter 6 constitutes the conceptual core of this thesis. There we develop a model that allows to acquire the structure of speech in terms of phones, syllables and words. We describe in detail the different subsystems of the architecture, and explain its processing pathways. The used evaluation metrics, the different kinds of evaluation scenarios and the obtained results are subsequently presented in chapter 7. There we validate that the proposed model reproduces some important aspects of infant speech development.

In chapter 8 we outline how our approach embeds into an embodied multi-modal learning and interaction framework. Furthermore we show how to conjoin our model with a developmentally plausible approach for speech imitation learning. The thesis concludes with a discussion 9, where we outline ideas for future extension, and link up the achieved experimental results with what is known from infant development as described in chapter 2.

CHAPTER 2

How language comes to children

If people ask the enclosing question of this chapter, they are commonly referring to two different but nevertheless highly interconnected issues. First, it is an open question how infants learn which sequences of speech sounds cohere to word forms. Second, it is an intensively debated issue how infants associate words and meaning. Both tasks – commonly referred as *lexical segmentation* and *vocabulary acquisition* – are somehow obvious from an adult's point of view but overwhelming complex from an infant's and – as we will see throughout this thesis – computational point of view.

Lexical segmentation needs to precede vocabulary acquisition at least to some part. This is because without words forms it is not possible to build and extend the mental lexicon. We might assume that words in speech are physically separable. However, when adults hear sentences in a unfamiliar language they are often unable to indicate word boundaries. What they perceive is a rather continuous stream of speech sounds. There are no gaps as in written text or special sound markers that could highlight word boundaries. The ability to segment speech into words is grounded in a deep unconscious representation of the language statistics.

As noted for instance in [Jus99b], it is clear that no single speech cue is solely sufficient to segment speech into words. Thus, to build a computational model of language acquisition we need to reveal which cues and what kind of language statistics underlay the process of word segmentation. Related, and even more important to us, are the bootstrapping processes that enable infants to start from nothing but basic speech reception, but to finally come up with the ability to convert acoustic utterances into words.

2.1 Child directed speech

Parents do not (want to) wait for their infants to master the problem of word segmentation before they start to interact. In contrast, most of them actively support this process by providing a special kind of tutoring speech. Different authors refer this as *Motherese* [Kit03], *Fatherese* [Bra08] or *child directed speech* (CDS) [Bat08]. Because we are in favor to the equality of rights, we will prefer the latter term throughout this thesis.

As summarized in [Bat08], CDS can be characterized by raised pitch, wider pitch range, exaggerated prosody, hyper-articulation, slower speech rate, and reduced linguistic complexity. Furthermore, it contains more frequent and longer pauses, is composed out of simpler and shorter utterances, is more repetitive, contains onomatopoeia and interjections more frequently, and conveys a reduced set of topics [Kit03]. CDS has been reported to convey more emotional information than adult directed speech. Parents seem to adopt to their infant's style of speaking, what extends their ability to control the infant's emotional arousal, and to focus its attention.

Although the role of CDS is still debated, there is evidence that 4-month-old infants prefer to listen to child directed speech than to adult-directed speech [Kit03, p. 7]. To dismiss the influence of gestures and face expression these results were obtained in a head-turning experiment that was conducted with audio tapes played to the infants.

The universality of CDS is challenged by the people like the Kaluli in New Guinea. There, adults address infants rarely, because they are not supposed to understand language. At the age of six to eight months they start to receive instructions [dBB99]. A similar scheme has been observed among the Kwara'ae of the Solomon Islands, where mothers speak to their infants indirectly. Frequently they speak about or even on behalf of their children by turning them toward the person who is being addressed [dBB99]. Further evidence that denies the mandatory character of CDS has been collected by observing people in Samoa and among some African Americans, which also do not use CDS in parent-infant interaction [Kit03]. The underlying purpose of such interaction patterns seems to integrate infants as early as possible into the social community.

Even under the assumption of CDS, the inner complexities of any spoken language are overwhelming complex. Infants need to process huge quantities of speech before they are start to segment speech successfully. Depending on type of caregiver the number of words addressed to infants within an hour differs considerably starting from fewer than 200 to over 3000 words. Within one hour, some parents spend more than 40 minutes interacting with their babies and some less than 15 minutes. Clearly, those difference add up, and the number words perceived until the age of 4 differs between 20 million and 50 millions [Har95].

2.2 Word segmentation

The simplest model for lexical acquisition is to assume CDS to contain isolated words at least as long as the infant has not acquired a sufficiently rich lexical model to generalize this knowledge to a continuous context [Dav01]. Synonymously, CDS could be supposed to comprise innately perceivable word boundaries like small silence periods that would reduce word segmentation to an innate ability. But such a model fails for several reasons. First, CDS does not consist of primarily isolated word utterances. Second, language-independent boundary cues are necessary to segment subsequent words, as anyone experiences when learning a new language.

However, isolated words seem to play an important role in bootstrapping more elaborate segmentation strategies. It has been shown that 9% of all CDS utterances in English are isolated words [Bre01]. Moreover, the frequency of words perceived in isolation strongly correlates with the timing of the infant learning this word [Gam05].

When do infants start to segment words from speech? Psychologists commonly address this question by using the head-turn preference procedure. Thereby infants are first familiarized with some auditory stimuli. Subsequently, they are exposed to patterns that are either new or familiar. During testing, the duration of the infant's head-turns towards the loudspeakers used to present the stimuli are measured. Significant increases of head-turn times for familiar patterns then provide evidence that infants have retained the knowledge of these stimuli.

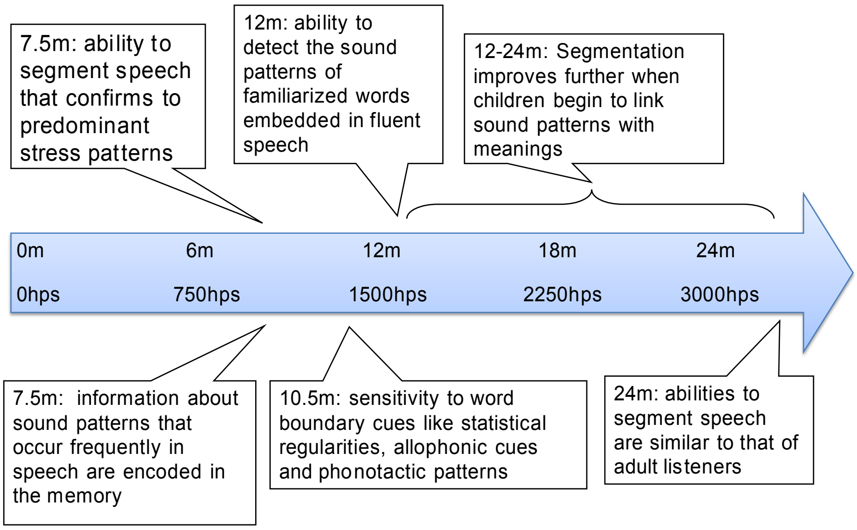

7.5m: ability to segment speech that confirms to predominant stress patterns
12m: ability to detect the sound patterns of familiarized words embedded in fluent speech
12-24m: Segmentation improves further when children begin to link sound patterns with meanings
0m
0hps
6m
750hps
12m
1500hps
18m
2250hps
24m
3000hps
7.5m: information about sound patterns that occur frequently in speech are encoded in the memory
10.5m: sensitivity to word boundary cues like statistical regularities, allophonic cues and phonotactic patterns
24m: abilities to segment speech are similar to that of adult listeners

Figure 2.1: Overview about word segmentation cue awareness in infants. *hps* denotes the estimated *h*ours of *p*erceived *s*peech until the corresponding age. [Jus99b]

Jusczyk et al. (cf. for a review [Jus99b]) used this technique to investigate when infants start to segment words from continuous speech. They showed in [Jus95] that 7.5 month-old infants that were familiarized with mono-syllabic words like `cup` listened significantly longer to sentences that contained these words than to utterances without any of the familiarized words.

Whatever learning strategies infants might employ, it is noted logically and developmentally, that infants will use language-independent segmentation cues prior to language-specific ones. This is especially of interest when drafting computational models of these processes. As synopsized in section 4, researchers often violated this finding by assuming non-innate language-dependent linguistic abilities to build models for speech segmentation learning.

2.2.1 Statistical Learning

By far the best known and best studied mechanism for speech segmentation is grounded in statistical regularities in the sound structure of a language. Whereas this idea can be applied on different levels of speech granularity, its commonly referred to the insight that syllables within a word tend to co-occur more frequently than those across word boundaries [Gam05]. This can be formalized by computing the *transitional probability* (TP) between adjacent syllables A and B

$$TP(A \rightarrow B) = \frac{P(AB)}{P(B)}. \tag{2.1}$$

Thereby, $P(AB)$ denotes the frequency of B following A, and $P(B)$ is the total frequency of B. By computing local minima of this function infants are believed to postulate word boundaries. Developmental evidence for this theory was provided by Saffran et al. in [Saf96]. Two-minute long sequences of synthetic speech composed of continuous repetitions of four different tri-syllabic words like `tibudo` or `pabiku` were presented to 8 month-old infants. These words were designed to contain an overlapping set of syllables. Infants were supposed to learn the order of syllables to segment utterances into words. And indeed they later preferred utterances that followed this scheme against utterances that were composed of words starting with the last syllable of one word followed by the first two syllables of another (e.g. `dopabi` or `kutibu`). Saffran and colleagues concluded that the infants had successfully learned the probabilistic syllabic transition model of the synthetic language.

Related to statistical learning is the idea that infants may obtain further word boundary cues by identifying utterance boundaries [Jus93b]. However, it is debated whether such cues are processed on syllable or phone level [Dav01] [Gam05].

In contrast to all other word segmentation strategies mentioned in this section, statistical learning seems to be the only language-independent word segmentation strategy [Gam05]. It relies solely on the identification of syllables within the speech input. Evidence for such a theory was provided by Thiessen and Saffran [Thi03] who reported 7-month-old infants to prefer statistical cues over metric cues when both cues are available. They infer that statistical learning may provide a set of seed words that allows bootstrapping of other language dependent segmentation strategies.

2.2.2 Metric segmentation

Stress is the defined as the relative emphasis given to a syllable within a word [Akm01]. Linguists further subdivide stress into primary and secondary stress to indicate the amount of emphasis given to particular syllables within a word. Stressed and non-stressed syllables are referred to as *strong* and *weak* respectively.

Reoccurring patterns of stressed and unstressed syllables are commonly referred as *metrics*. Metric patterns provide another cue that allows humans to segment speech into words. The underlying principle is based on stress patterns that indicate the beginning, center or end of a word. For instance, 90% of all English words are stress initial [Cut87]. This gave rise to the assumption that the young learners treat stressed syllables as beginning of words [Cut88]. Although similar metric cues can be applied to other languages, metric segmentation is not suited for languages without a clear dominant stress pattern (like French). It is not yet clear how metric segmentation is learned by infants because it must be preceded by other segmentation strategies to provide a sufficiently large training sample. However, this renders metric segmentation unlikely to be a central word segmentation learning principle, as it leaves open how infants obtain these seed words.

Another prominent and intuitive metric segmentation principle is related to the *unique stress constraint* (USC) [Gam05]. It states that *a word can bear at most one primary stress*. Importantly, this principle does not rely on a particular manifestation of stress. It is believed by some researchers to be an innate [1] phonological language independent constraint on word boundaries. Therefore, it can be considered to equip infants with an initial indication mechanism that allows to identify isolated words as such. However, USC becomes operative not before the young learners have acquired a stress model of their parent's language. And because stress manifests itself in many different ways, it is not clear how infants can apply this principle unconditionally.

[1] Gambell and Yang infer that USC is innate because of an – in our opinion arguable reasoning: As they've shown experimentally in [Gam05] statistical learning alone does not result in a sufficiently stable set of seed words. Therefore USC cannot be learned in a data-driven manner. Because they assume statistical learning to be the only language-independent way to learn word segmentation, USC must be assumed to be innate. However, in our opinion other language-independent principles for word boundary detection may have not yet been discovered. We are not aware of any developmental evidence that infants do *not* rely on additional – not yet revealed – segmentation cues. In our opinion it is more likely that seed word extraction relies on statistical learning complemented by other language independent principles that allow to bootstrap USC in a similar manner as other cues for lexical learning.

Furthermore, even under the assumption that infants are able to detect a set of set of seed words, it is not clear how they identify metric patterns in speech. This requires the ability to reveal how stress manifests in a particular language. This is commonly believed to be a subset of the many different stress types like *quantitative* (length), *dynamic* (loudness), *qualitative* (fuller vowels) or pitch stress [Akm01]. For instance, stressed syllables in English have higher pitch, longer duration, and typically fuller vowels than unstressed syllables, as well as being dynamically louder. Clearly, respective subsets have to be identified for any particular tutoring language prior to the learning of reoccurring stress patterns.

There is a considerable body of evidence that supports the idea of metric segmentation. For instance, 7.5-month-old infants do better at recognizing English utterances that contain strong/weak words that those with weak/strong patterns. 9-month-old English infants prefer strong/weak patterns in words over those with the weak/strong ones [Jus93a]. According to [Kit03, p. 24] English-learning infants may learn the rhythmic cues to word onsets and the metrical segmentation strategy (MSS) strategy from their experience of listening to isolated words, in particular, English first names. According to Elena Lieven 45% of mothers utterances start with one of 17 words [2], which may facilitate the extraction of an initial set of seed words because of the high number of repetitions in the beginning of utterances.

Interestingly, it was also shown that infants are overconfident with respect to their metric segmentation abilities. In [Jus99c] Jusczyk and colleagues found 7.5-month-olds to treat `taris` in `(guitar is)` as a word. They argue that this finding can be explained by the fact that `tar` is a strong syllable which indicates the beginning of a new word according to the dominant strong/weak pattern of the English language.

2.2.3 The principle of subtraction

With respect to the framework we aim to develop in this thesis, we are especially concerned with the processing direction in which infants acquire the lexical structure of language: Do they rely on bottom up processing or it is rather a top-down directed process that allows to reveal new words? Linguists have suggested that infants may lack the resources for top-down segmentation and instead rely exclusively on bottom-up cues in the speech stream [Cut96] [Jus97a].

However, the findings of Bortfeld and colleagues partially contradict this view: As reported in [Bor05] infants do rely on top-down feedback within one level of speech granularity. They have shown that infants as young as 8 months prefer word pairs that were composed of a familiar word (like `mommy`) and unknown test words to pair solely composed from unknown test words. This makes the authors to infer that infants can already exploit highly familiar words to segment and recognize adjoining, previously unfamiliar words from fluent speech.

Whereas some words may appear as isolated seed words for word structure acquisition, most words will not appear without a continuous speech context. Nevertheless young learners seem to learn new words with breathtaking speed as soon as they have acquired an initial set of seed words. It has been estimated that infants acquire about 9 words per day from the age of 1.5 to 6 years [Car78]. This process is commonly referred to as *lexical explosion* [Akm01] [Jus97a] [Goo98].

[2] Keynote talk at EELC, Rome, 2007

Although the principles behind these findings are still far from being understood completely, many researchers consider the *principle of subtraction* to play an important role in this process. Its basic idea is that an already acquired representation can be used to compute a kind of residual given a sensory stimulus by suppressing or removing already represented parts [Pet83]. For instance, let the young learner to have acquired the words `house` and `the`. A tutoring utterance like `(the red house)` would result in the residual `red`. Especially by assuming CDS to contain a set of isolated seed words this becomes a logically clear and consistent model for lexical acquisition.

Beside the above mentioned study of Bortfeld, more evidence for such a residual-based acquisition principle was provided by Jusczyk and Hohne in [Jus97b]. They showed that 8 month-olds listened significantly longer to stories with previously familiarized words embedded, than to utterances containing foil words. Remarkably, this held even weeks later after the training phase. The used familiarization words like `python`, `vine` or `peccaries` were carefully selected to play no role in a common infant's world.

In addition, English speech corpora analyses in [Gre98] revealed that the ten most frequent words account for approximately 25% of all lexical instances. One hundred words account for 66% of all corpus words. So even if statistical or metrical cues are not reliable for rare words, subtraction seems to provide a powerful principle to extract novel words from a continuous context. It is clear that by applying the principle of subtraction a lexical model will grow exponentially in the number of words until the complete lexical structure of the language has been covered.

2.2.4 Allophonic and articulatory cues

Logically less feasible and more subtle stimuli that provide information about word boundaries are *allophonic* and *articulatory* cues. The former are the result of context-dependent variations of articulation. For instance the allophone /t/ is pronounced aspirated at the end a word (e.g. ”cat”) but non-aspirated at its beginning (e.g. ”table”). Related but more flexible with respect to the context are articulatory cues. These represent the result of co-articulation between adjacent phones, that varies as a complex function of their positions within or across syllable/word boundaries [Lad93].

It is rather unclear how infants acquire knowledge about such those cues. Like for metric segmentation a set of already segmentable seed words is at least necessary to infer the properties of these cues. Supplementary, to some extent infants may infer their knowledge about co-articulation effects from their own organization of speech production [Bro92]. However, a purely innate representation seems unlikely because of evidence presented in [Jus99a]. There, Jusczyk and colleagues showed that infants seem to be unaware of allophonic variations between ”nitrates” and ”night rates” that would enable them to find the word boundaries in the latter.

2.3 Syllable segmentation

Whereas all above mentioned word segmentation principles considerably differ in their nature, they are sharing a common concept: The idea that *words are composed of syllables* [Gam05, p. 22]. So the first step when thinking about models for lexical acquisition is to investigate the structure of

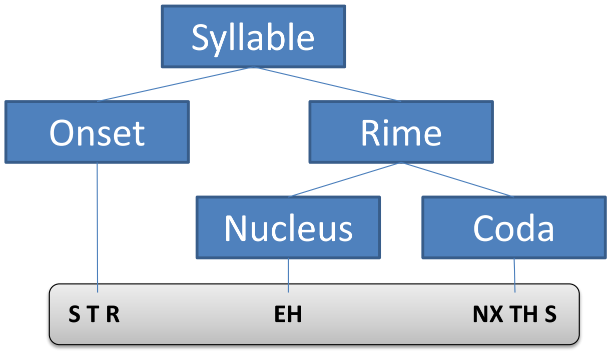


Figure 2.2: The linguistic definition of a syllable.

syllables.

There is a considerable disagreement in the literature about how syllables should be defined. Syllables are a linguistically *slippery* concept that remains difficult to be pinned down precisely while maintaining their notion as units with an intuitive plausibility [Hua01].

A fuzzy, but broadly accepted view is that syllables are consonants enclosing a center vowel. More specifically, linguists define syllables to consist of an *onset* and a *rime* [Gre98]. As depicted in figure 2.2 the latter can be further subdivided into a *nucleus* and an optional *coda*. Whereas onset and coda consist of one or several consonants, the nucleus consists of a sequence of vowels.

Such a notion of syllables is challenged by languages that allow long strings of consonants without any intervening vowel or sonorant like Nuxlk [dBB99]. Therefore we follow a more generic definition, that *syllables* are *phonotactically constrained series of phones* [Gam05]. This definition does not make any assumption on the syllabic structure but a framing into chunks of phones according to phonotactics of the language. *Phonotactics* refer to structural restrictions on what makes a well-formed syllable in a particular language. For instance, only certain consonant clusters can serve as onsets for English syllables: ”clight” or ”zight” are not actual English words, but they are valid syllables with respect to the phonotactic constraints of the English language. In contrast, ”nkight” or ”dnight” are not well-formed according to phonotactic constraints of modern English [Jus99c]. Such phone sequences are permissible only at certain locations.

The linguistic definition of a syllable needs to be treated with special care when dealing with acoustic speech instead of symbolic phone sequences (as preferred by linguists). Syllables vary markedly from their canonical structure, when being realized depending on the speaking rate. That is, phones are modulated, dropped or even substituted. Although such effects seem to appear randomly on a phonetic level, they can be structured when a syllable framing of speech is assumed. Greenberg [Gre98] condensed such variations into a small set of principles.

1. Syllable onsets are generally preserved.
2. The nucleus often deviates from its canonical form.
3. Coda elements are often deleted.
4. The amount of co-articulation effects is inversely correlated with the information valence of a word in an utterance.

Brief, the likelihood of canonical expression decreases within the syllable. It has been suggested that bio-mechanical constraints imposed by the vocal tract cause these effects. However, this is arguable because of the fact that onsets of subsequent syllables should suffer from the same constraints as the respective preceding coda elements. More plausible are evolutionary selection forces. The auditory system is more sensitive to onsets, because they can serve as alert signals and are thus considered to be more informative, than medial or terminal syllable constituents.

Even under the assumption of the more perceptually motivated syllable definition, the question remains how infants learn the phonotactic constraints from the tutoring speech. There are hints that speech acquisition might already start before birth. Kit [Kit03, p. 4] notes, that newborns are already particularly sensitive to syllabic structures. This idea has been supported by findings of [Meh88a], that infants are able to discriminate speech in their mother tongues from speech in other languages based on specific prosodic patterns.

However, such findings always depend on the definition of a syllable. A popular assumption is that infants are born with nascent structure-seeking mechanisms to discover distributional patterns in the speech input, promoted by innately specified structural constraints (cf. [Kit03] for a review). This has been supported by some amount in neonate perception studies of [Jus97a], who concluded that syllabic structure properties may be innate. However, there is no agreed model about what such constraints could look like. For instance, an innate syllable length could be constrained by the frequency of chewing-patterns, which occur on a similar time-scales like syllables. More language dependent constraints are questionable, because this would require them to be genetically encoded. Even more high level constraints, as Chomsky's ideas about an innate universal grammar [Cho65] to support language acquisition, have been already rebutted.

In any case innate constraints do at best play an initializing or supporting role. This is because the infant's sensitivity to different structural properties of speech input changes with their age. As summarized in figure 2.1 infants require up to 12 months to become aware of some cues that are believed to be mandatory for lexical learning. Therefore, to achieve a better understanding when infants are sensitive to which syllable segmentation cues, we highlight the most important principles including references to psychological studies in the following subsections.

2.3.1 Sonority sequencing principle

Even without having developed a complete model of phonotactics, researchers have condensed out some phonotactic principles. A very important one that applies to a wide variety of languages is the *sonority sequencing principle*. Its idea is that syllable nuclei correlate with peaks in sonority. Thereby, *sonority* refers to high-amplitude, periodic and often vocalic sections of the speech waveform. According to this principle, sonority increases monotonously up to a peak level and than decays monotonously until the offset of the syllable. By imposing a scale of sonority on all speech sounds, consonants can be ranked in terms of sonority. Decreasingly ordered this gives the following scale: glides, liquids, nasals, affricates and fricatives and stops [Akm01] [Hua01, p.51].

This scale along with the sonority principle can be straight forward applied to parse a speech sound sequence into syllables. Boundaries need to become inserted at local sonority minima. This process is referred to as *syllabification*. However, due to the nature of speech this cannot applied unambiguously in all cases. Syllabification must be complemented by other phonotactic constraints on the constituents of a syllable. Furthermore, higher-order consideration of word structure may need to complement sonority-based syllabification.

2.3.2 Maximum onset principle

Another phonotactic principle that is believed to contribute to syllabification is *maximum onset*. It states that the onset before each nucleus can be extended as long as it is valid with respect to the phonotactic model of the language. It has been successfully implemented to syllabify phonetic speech corpora. However, it is still debated whether the maximum onset principle is bound to certain languages or denotes a general principle. Clearly, its power strongly depends on the language under consideration. Whereas it applies well to many Asian languages because of their clear $C-V$ structure, it is less suited for European languages like German or French.

2.3.3 Phonotactic learning

How do infants become aware of phonotactic constraints without observing isolated syllables? By assuming them to be innately able to detect utterance boundaries, they may bootstrap their phonotactic model by simply paying attention to initial and final parts of perceived utterances. Every utterance necessarily starts with a syllable and ends with a syllable, so infants may infer phonotactic constraints of their language given a sufficiently amount of tutoring from utterance boundaries only.

Such an approach may be infeasible for certain languages. Because of the combinatorial explosion of possible phone combinations even with a syllabic structure comprising just a few phones, an enormous amount of speech would be necessary to acquire a sufficiently stable phonotactic model. However, for languages like English only a few dozen possible syllable onsets and codas are valid which make it reasonable that infants learn phonotactics from the utterance boundaries only [Gam05].

Whereas most phonotactic principles need to be derived from the language being learned, some authors [Gam05] argue that syllable segmentation mechanisms must be complemented by what appear to be innate constraints on phonological structures. This is because the articulatory system imposes certain restrictions on sound sequences that can be produced. Nevertheless it remains unclear to which amount this applies due to the fact that for each phonotactic principle there are counter example languages to which it does not apply.

It has been suggested that phonotactics also play a direct role in word segmentation. However, such an impact is commonly believed to be less direct [Gam05, p. 6]. Instead words are assumed to assemble from syllables. This indirectly links words to phonotactics and thus integrates syllabification and lexical segmentation. This indicates also the order of acquisition: phonotactics need to be acquired prior to word segmentation strategies.

2.4 Phone segmentation

Because phones from the basis of phonotactics, a mandatory first step to learn the syllable structure is to bootstrap a phone representation. Hereby, we consider *phones* to be speech segments that possess distinct physical and perceptual properties. Phones are basic speech sounds without any relation to meaning. In contrast, *phonemes*[3] are a linguistic concept and refer to minimal meaningful sounds.

At birth, infants are sensible to all sound contrasts in all spoken languages. However, they start to lose sensitivity to phonemic contrasts that are not relevant for their native language [Jus97b] and gradually adapt to the phonology of their parent's tongue. This indicates that infants could learn phones in a generalizing manner, starting from fine-grained perceptual abilities which converge against a more general representation linked to their parental language.

Not much is known about how phone learning could be organized in infants. This is because it is hard to setup experiments to assess the infant's recognition or learning ability on such a subsyllabic level. Nevertheless some studies have investigated the discriminative abilities of infants for different languages. For instance, infants as young as few days are already aware of differences between their mothers language and other ones [Meh88b]. Furthermore, they prefer to listen to natural language than to other auditory inputs [Gle81]. This strengthens the idea that speech acquisition may begin to some part already before birth.

2.5 Vocabulary acquisition

Without meaning acquired units are just structural elements. Following [Kit03] we refer by *meaning* to the mental representation of concepts in our mind referring to objects and their properties in the real world. The meaning of speech has to be inferred from a complex multi-modal perceptual context. Although the development of speech segmentation has to precede the association of meaning to series of observed speech units (see [Dav01, p. 4, 20, 32]) both processes are likely to complement each other to cohere into entities of the mental lexicon.

It is still an unsolved question how infants ground acquired lexical units because there is an infinite number of referees that a word could denote [Qui60]. Given a toy duck and the word `duck`, the latter could refer to name, color, size, parts of the object or a completely unrelated other property in the surrounding scene. Some researchers have suggested that infants place constraints about referents of words. For instance, they may expect words to refer to objects and taxonomic categories. Furthermore, they assume every two words to contrast in meaning (cf. [Goo98] for a review).

Another cue that may support vocabulary acquisition is the *morpho-syntactic* context in which a word appears. For instance, a determiner (like *the*) occurring before a sentence-final term strongly indicates a novel word to be a countable noun [Goo98]. Although this requires an already sophisticated understanding of syntax and morphology, infants younger than 2 years were able to discriminate between proper names and category names because of this principle [Kat74]. Similar mechanisms were shown to be employed by infants to determine whether novel terms refer to nouns or adjectives [Tay88]. Furthermore, infants seem to be biased to assume new words to refer to whole bounded objects rather than their properties or parts [Wax96].

[3] Many works cited throughout this thesis use the term *phoneme* to denote short units of speech that possess distinct perceptual properties. Clearly this is misleading and incorrect given the linguistic background of the term, and clashes with the definition of *phones*. *Phonemes* refer to minimal meaningful sounds not to perceptual units. However, to keep our citations continent with the referred works, we keep the arguable use *phoneme* in such cases.

Whereas the underlying learning mechanisms remain unclear, psychologists have investigated at which age infants start to link words and meaning. Commonly researchers rely on the *preferential looking* technique, where infants are presented with a set of objects and a single word describing one of them. Then, their tendency to fixate a particular object is assumed to give evidence about the infants understanding of the word. Using this technique Thomas and colleagues [Tho81] have shown word-object relationships to be present 13-month-old infants. However, using an improved setup which compared looking preferences for appropriate against inappropriate objects infants were reported to understand words not before 15 months [Gol87].

According to [Roy00], infants already benefit from multi-modal input during early phases of speech acquisition. But what could be the underlying principles that ease this grounding process of speech symbols? First, integrated multi-modal sensory processing tends to outperform unimodal processing. For instance, it was shown that combined auditory and visual stimuli give reaction times are significantly better than for unimodal ones [Rom07]. Second, multi-modal cues have been reported to interfere with speech abilities [Str35], which may indicate that the brain imposes some kind of top-down control on speech perception processes in presence of non-speech stimuli. Finally, 2-year old infants were shown to infer the meaning of novel nouns using the semantic context, and to retain those meanings a day later [Goo98].

Another principle referred to as *lexical contrast* applies when familiar words are paired with unfamiliar ones of the same category (e.g. "give me the red cup, and not the green one, please"). Evidence that 3-year-olds can use lexical contrast to infer the meaning of novel words was presented in [Au90]. However, counter-evidence [Hei87] indicated that lexical contrast seems to be not mandatory to succeed in learning new words.

CHAPTER 3

Pattern recognition background

The approach developed within this thesis touches many areas of pattern recognition like speech recognition, clustering, unsupervised learning, graph theory and optimization. In this chapter we outline the most relevant approaches and computational frameworks that defined the starting points of our research.

3.1 Clustering

Clustering refers to a process that reveals a description of a data set in terms of groups of samples that possess strong internal similarities. Formally, a clustering procedure for a data set $\mathcal{D} = x_1, ..., x_n$ is defined by

1. A *similarity measure* $d(x, x')$ that implies a natural grouping of sample elements with respect to the application. A common choice is the *Minkowski metric*

$$d_M(x, x') = \left(\sum_{k=1}^{d} |x_k - x'_k|^q \right)^{1/q} \tag{3.1}$$

   that can be applied to any data-set drawn from a linear d-dimensional vector space. $q \geq 1$ denotes an arbitrarily selectable parameter. For instance $q = 2$ gives the most common *Euclidean* distance.

2. A *criterion function* J to be optimized. Probably most popular is the *Least-Squares criterion* that is defined by

$$J_{LS} = \sum_{i=1}^{c} \sum_{x \in \mathcal{D}_i} ||x - m_i||^2 \tag{3.2}$$

   Hereby, c denotes the number of clusters, $\mathcal{D}_i$ the ith cluster and m_i its mean. As many criterion functions, J_{LS} is designed to approximate the scatter of the clustering.

3. An actual *clustering procedure*. Formally, to determine an optimal clustering all $\frac{c^n}{n!}$ different possible partitions of n data points into c clusters need to be considered. This is combinatorially not feasible even for mid-size problems. Thus, a clustering procedure has to be selected. Its choice depends on various application-specific factors: A fixed number of clusters versus a dynamically changing set of clusters, batch- versus online-clustering, hierarchical versus partitioning methods, the properties of the used distance measure, or the amount of supervision available to the system.

```
Require: $k \in N : k < |\mathcal{D}|$
  Initialize the cluster centers $\mu_1, ...\mu_k$
  while convergence condition has not met do
    Assign each $x_i$ to the nearest cluster
    Recompute $\mu_1, ...\mu_k$
  end while
```

Figure 3.1: k-Means clustering. Besides being popular because of speed and simplicity, its main drawback is that the number of clusters has to be defined heuristically. As most clustering approaches, it implements an iterative optimization procedure to reveal a local-minimum solution with respect to the chosen criterion function.

Because of the *no free lunch theorem* there is no general best solution without considering a particular problem [Dud00, chp. 9]. This theorem states, that no matter how clever we are in selecting solutions for all three clustering issues, the resulting solution will not even outperform random guessing unless we restrict the class of problems.

3.1.1 Incremental clustering

An actual clustering procedure crucially depends on whether the sample is present as a whole or accumulates over time. For many applications the complete sample is present on startup. This allows to use methods like *k-means* clustering (cf. fig. 3.1) or *hierarchical clustering*. Instead of revealing a single clustering, the latter defines a sequence of partitions with the property that two items that are in one cluster at level l are also in the same cluster for levels $l' > l$. With respect to data-driven learning, this is beneficial because such a sequence allows to define a metric on the data points, but requires higher computational efforts (cf. [Dud00, sec. 10.9.4]).

With respect to infant inspired speech development it is however clear, that tutoring speech of the infant's caregiver at best accumulates incrementally. This restricts the range of applicable clustering methods to those approaches that reveal a clustering incrementally without initial assumptions about the number of clusters. However, as we will discuss later, some constraints can – and need to – be considered to be innately available to the infant.

Even if often confused or dismissed in the literature, we attempt to make a clear distinction between incremental and online learning. We consider methods to be *incremental* if the portion of a sample available to the clustering procedure accumulates over time. In contrast, we refer approaches as *online*, if they process each data sample as it occurs in time *without* having a possibly unlimited memory. In this sense online methods can be regarded as a memory-less instances of incremental clustering approaches, which by design disregard new sample elements as they have processed them.

Such a rejection of data seems unnecessary at a first glance, but it makes sense because of two reasons: First, the amount of speech necessary to bootstrap a language representation is extremely huge as highlighted in chapter 2. This might make a computational model with an unlimited history hard to realize on a computer with finite resources. Second and more important, infants are unlikely to keep all training data in mind as it, but rather develop an abstract speech representation.

Incremental approaches have to deal with a problem known as the *stability-plasticity dilemma* [Gro88]: A system has to be sufficiently adaptive to learn from new data. On the other hand side, most recently processed samples may cause a major reorganization of the representation, so that

```
Require: $\theta > 0$, $\eta > 0$
  while Has new $x$ do
    $j = \arg\min_i ||x - \mu_i||$
    if $||x - \mu_j|| < \theta$ then
      $\mu_j \leftarrow (1 - \eta)\mu_j + \eta x$ {Adapt winner with current input}
    end if
  end while
```

Figure 3.2: Leader-Follower Clustering

previously acquired knowledge might get lost. It is up to the system designer to find a balance between both depending on the structure of the data to be clustered.

One approach to overcome the stability-plasticity dilemma is the use of local optimization criteria to confine the influence of a sample to the sub-set of clusters it is related to. Often this idea is implemented as *competitive learning* where only the best matching set of clusters is adapted given a data.

Beside plasticity issues, the central problem of clustering is to choose an appropriate number of clusters to be estimated. One approach is to solve the problem for many different values of c. If the score function of the used clustering criterion with the number of clusters as independent variable indicates a large gap for increasing values of c, the natural clustering is supposed to be found. Because of its characteristic shape, this function is referred to as *elbow* function. However, for real world problems a clear "elbow" may be observed occasionally only.

A second approach to determine the number of clusters in a data-driven manner that especially targets online-learning is to use a *similarity-* (or inversely formulated *novelty-*) threshold θ to decide when to create new clusters. An adapted k-means procedure that implements this principle is outlined in algorithm 3.2. Compared with algorithm 3.1 this implementation requires only linear time. However, it leaves open of how to select the novelty threshold θ and the learning rate η that balances stability against plasticity during learning. Of special concern in this work are more elaborate methods for *novelty detection*, which we use to refer to the identification of new patterns that a machine learning system is not aware of in prior training [Mar03]. In particular we focus on methods to determine the novelty of speech segments in chapter 6.

A popular extension to Leader-Follower clustering is referred to as *growing neural gas* [Fri94]: In addition to leader-follower clustering, it implements edges used to link newly created clusters and the n-best matching ones. Each of these links is associated with an *age*-parameter, that is initially set to 0. After each update step any such edge age becomes incremented by 1 and edges that exceed a certain age-threshold become removed. Finally, edge-less nodes are deleted, to ensure a representation that fits to the data without focusing on spurious or transient noise patterns.

3.1.2 Self-organizing neural nets

An approach for online-learning that is related to Leader-Follower clustering (cf. algorithm 3.2) has been presented in [Koh89]. By assuming clusters to be topologically arranged on an n- but in practice almost always 2-dimensional grid, not only the best matching cluster but also its neighboring clusters are updated. Thereby the shift towards the current input decays with the topological distance between a cluster to the best matching one. This is usually implemented as a

```
Require: $\eta > 0$, $\mu_{1:c}$ randomly initialized
  while Has new $x$ do
    $j = \arg\min_i \mu_i x$
    for all $\mu_i$ do
      $\mu_i \leftarrow (1-\eta)\mu_j + \eta\wp(\mu_j, \mu_i)x$
    end for
  end while
```

Figure 3.3: Self-organizing neural net

window function $\wp(\mu, \mu')$ that is 1 for $\mu = \mu'$ and decays according to the topological distance of μ and μ'. Another important difference to the above mentioned clustering approaches is the used similarity measure, which is chosen to be the dot product between cluster center and the current input. This is motivated by idea that clusters are modeling *neurons* which weight the elements of their multidimensional input to produce a neural activation.

Because of this biologically inspired processing this approach is called *self-organizing neural net*. It has been shown that such a network architecture is able to preserve topological orderings under various conditions [Koh88].

3.2 Probability density estimation

A central problem of machine learning is the estimation probability density functions $p(x|c)$ given a class c. Let $\mathcal{D} = x_1, ..., x_T$ a data set. Without any prior knowledge about such a density, the task is, to find a solution for

$$p(D|c) \rightarrow \max! \tag{3.3}$$

3.2.1 Parametric approaches

Problem 3.3 becomes more tractable when the sample $\mathcal{D}$ is assumed to be drawn from a generative *parametric* model. This is a valid premise when the central limit theorem applies or the structure of the system under consideration is well known. Then the problem simplifies to the estimation of $p(\mathcal{D}|\theta)$, where θ is a parameter vector containing all parameters of the assumed parametric model.

The *likelihood* of θ with respect to the sample $\mathcal{D}$ is calculated by

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{T} = p(x_k|\theta) \tag{3.4}$$

Under the assumption that the elements of $\mathcal{D}$ have been drawn independently, a solution of problem 3.3 is given by the *maximum likelihood estimate* of θ, which is by definition the value $\hat{\theta}$ that maximizes equation 3.4. Concretely, this is the instance of the assumed probabilistic model that fits best the sample $\mathcal{D}$.

For analytic reasons the logarithm of eq. 3.4, which is referred to as *log-likelihood*, $l(\theta) = \ln p(\mathcal{D}|\theta)$ is used to compute $\hat{\theta}$ via

$$\hat{\theta} = \arg\max_{\theta} l(\theta) \tag{3.5}$$

$$\nabla_\theta l = \sum_{k=1}^{T} \nabla_\theta p(x_k|\theta) \tag{3.6}$$

$$\nabla_\theta l = 0 \tag{3.7}$$

Because solutions of eq. 3.7 do not hold only for the global maximum to be found, higher order conditions must be checked additionally. It is important to note that $\hat{\theta}$ defines just a function for the parameters of $p(\mathcal{D}|\theta)$ which depends on the sample $\mathcal{D}$. Only for $n \to \infty$ it will converge against the true value of the generating function [Dud00, chp. 3].

Example. Let a sample $\mathcal{D}$ be drawn from a univariate Gaussian. The log-likelihood of a single sample element is calculated by

$$l(\theta) = l([\mu, \sigma^2]) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x - \mu)^2 \tag{3.8}$$

The maximum likelihood solution for μ and σ can be obtained according to equation 3.7. Thus, setting the partial derivatives with respect to $\theta = [\mu, \sigma^2]$ to zero gives the desired solution:

$$\hat{\mu} = \frac{1}{n} \sum_{k=1}^{T} x_k \tag{3.9}$$

which is the sample mean, and the (biased) sample co-variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{T} (x_k - \hat{\mu})(x_k - \hat{\mu})^t \tag{3.10}$$

Further discussion of ML-estimates concerning their distribution and asymptotic properties like bias, efficiency or consistence have been presented in [Dud00].

3.2.2 MAP

As discussed above in section 3.1.1 computational models for speech acquisition need to implement incremental rather than batch clustering techniques, because training data will be observed incrementally in interaction with the system's caregiver. Whereas ML-estimation of clusters is applicable and useful for many applications, it is not applicable the context of this thesis.

By following a more Bayesian approach to treat the parameter vector θ as a random variable, the posterior density $p(x|\mathcal{D})$ can be estimated incrementally as follows. We assume the parametric form of the density $p(x|\theta)$ to be known without knowing the value of its parameterization θ exactly, the initial knowledge about θ to be contained in a known a priori density $p(\theta)$, and the remaining knowledge about θ to be contained in $\mathcal{D}$. Than the formal solution is given by

$$p(x|\mathcal{D}) = \int p(x|\theta)p(\theta|\mathcal{D})d\theta \tag{3.11}$$

which can be transformed into an actual computational solution by applying the Bayes' formula

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta} \tag{3.12}$$

and by dissolving the independence assumption

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{T} p(x_k|\theta) \tag{3.13}$$

The effect of observing additional samples is to sharpen the posteriori density function that is initially dominated by the prior knowledge about θ as summarized in the prior $p(\theta)$, causing it to peak near the true value of θ. Thus, the optimal solution θ_{MAP} is obtained by simply picking the mode of the posterior

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta) \tag{3.14}$$

This model estimation approach is referred to as *Bayesian learning.* It reduces to ML-estimation when the prior is $p(\theta)$ is a uniform distribution.

Three main issues need to be addressed when implementing a MAP-estimation scheme [Gau94]. First, the choice of the prior distribution family. Second, the parameterization of the prior density. And finally, the actual calculation of the MAP-estimate. All three problems are related to each other. As pointed out in [Gau94] a computationally feasible MAP-estimator can be chosen if $p(x|\theta) \propto f_X$. Hereby, f_X denotes the exponential family. This is because such p.d.f.'s always possess a sufficient statistic of fixed dimension with respect to θ, which allows to select a prior in a way that the resulting posterior $p(\theta|x)$ simplifies to the same algebraic form as $p(\theta)$. A *statistic* $\mathcal{S}$ denotes any quantity derived from a sample $\mathcal{D}$. It is said to be *sufficient* with respect to a distribution parameter θ, when the posterior $p(\mathcal{D}|\theta, \mathcal{S})$ is independent of $\mathcal{D}$. Brief, a sufficient statistic $\mathcal{S}$ captures all information with respect to θ, which means that $\mathcal{S}$ allows to reduce the information contained in $\mathcal{D}$ to a small set of values [Dud00].

3.2.3 Non-parametric approaches

If it is not possible or reasonable to make assumption about a data set's underlying generative modes *non-parametric density estimation* techniques may be applied instead of parametric ones. In contrast to the latter, the former do not rely on statistics like mean or variance. Instead, the sample as it is used to setup a functional probabilistic density model. A powerful implementation of this idea are *Parzen windows*, which approximate a data-set with a density function obtained by summing a weighted set of kernel functions centered around the sample elements.

$$p_T(x) = \frac{1}{T}\sum_{i=1}^{T} \phi\left(\frac{x - x_i}{h}\right) \tag{3.15}$$

Thereby, ϕ denotes a non-negative *kernel* function with a *kernel-width* h, which is centered at x_i. Under the assumption that the kernels itself are probability densities a normalization with respect to the sample size is sufficient. Otherwise each kernel needs to be normalized by the kernel's volume. Typically Gaussian, triangular or rectangular kernels are used.

Parzen window estimates can be shown to converge against the true density for $T \rightarrow \infty$. Another important property of the Parzen windows is the possibility to add more kernels dynamically. Especially with respect to the aim of this thesis, this makes such models attractive for incremental learning. However, the computational costs when evaluating a non-parametric model are magnitudes higher compared to a parametric one, which may outweigh their virtues when dealing with systems for real-time speech processing.

3.3 Information theory basics

One of the central concepts when dealing with unsupervised learning, is the notion of *novelty.* Information theory provides powerful tools to quantify novelty within a probabilistic framework.

The central measure is *entropy* $\mathcal{H}(X)$ of a random variable X, which quantifies the uncertainty associated to X [Dud00]. It is calculated by

$$\mathcal{H}(X) = -\sum_{k=1}^{K} P(x_k) \log P(x_k) \tag{3.16}$$

With respect to the aim of this thesis it will be important to superimpose a metric over a set of probabilistic speech unit models. One such metric can be defined based on the *Kullback-Leibler distance* (also known as *trans-information* or *relative entropy*) that approximates the distance from a target distribution p to a test distribution q by

$$D_{kl}(p|q) = \int p(x) \log \frac{p(x)}{q(x)} \tag{3.17}$$

Whereas $D_{kl}(p|q)$ itself is non-symmetric (and therefore not a metric) it can be shown that

$$\hat{D}_{p,q} = \frac{D_{kl}(p|q) + D_{kl}(q|p)}{2} \tag{3.18}$$

defines a metric [Jua85]. Whereas this makes $\hat{D}_{p,q}$ applicable to clustering tasks, it comes along with a high computational burden, that it to sample p and q many times. However, fast approximation schemes have been proposed that allow to calculate $\hat{D}_{p,q}$ directly at least for some types of models like mixture of Gaussians [Che05]. Unfortunately, for complex multivariate mixture density HMMs no such approximations schemes have been proposed yet. For those D_{kl} can be calculated only by approximating the integral by a sum over a large set of randomly sampled observation sequences.

Related to D_{kl} is *mutual information* $\mathcal{I}(X, Y)$ which measures the mutual dependence between two random variables X and Y. Intuitively, mutual information measures the information that X and Y share: it measures how much knowing one of these variables reduces our uncertainty about the other. It is defined as

$$\mathcal{I}(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{3.19}$$

Thereby, p(x,y) denotes the joint probability of x and y. If both are independent from each other, $\mathcal{I}(X, Y) = 0$.

3.4 HMMs

Any computational model that attempts to grasp the nature of speech must take the time evolving nature of speech into account. Even if there is a great variety of such models available, the vast majority of systems is based on Hidden Markov Models (HMM).

A Hidden Markov Model is a probabilistic model of the joint probability of a collection of random variables $\{X_1, ..., X_T, S_1, ...S_T\}$ which are supposed to satisfy two independence assumptions. The X_t refer to either discrete or continuous observations, whereas S_t are non-observable *hidden* and discrete. The hidden process obeys the *Markov*-property, that the hidden state a time t solely depends on its state at $t-1$:

$$P(S_t|S_1, S_2, ...S_{t-1}) = P(S_t|S_{t-1}) \tag{3.20}$$

Second, the observation at time t depends only the current value S_t.

$$P(X_t|X_1, X_2, ...X_{t-1}, S_1...S_t) = P(X_t|S_t) \tag{3.21}$$

Whereas the sequence of X is observable, the underlying state sequence S remains hidden. Such processes are considered for finite sequences only, so the initial state of such a model needs to be specified.

A 1st order Hidden-Markov model $\lambda = \{\pi, \mathbf{A}, \mathbf{B}\}$ with N states is fully characterized by

- a finite set of states $s_{1:N}$, commonly referred to only by their indices.
- a matrix A that specifies transition probabilities between these states

$$\mathbf{A} = \{a_{ij}|a_{ij} = P(S_t = j|P_{t-1} = i)\} \tag{3.22}$$

- a vector π of initial state probabilities.

$$\pi = \{\pi_i|\pi = P(S_1 = i)\} \tag{3.23}$$

- state-dependent emission probability density functions

$$\begin{aligned} \mathbf{B} &= \{b_j(x)\}_{j=1:N} && (3.24) \\ b_j(x) &= p(x|S_t = j) && (3.25) \end{aligned}$$

  Depending on the type of observations the b_js are either discrete densities over a finite probability space $\{o_1, o_2, ...o_M\}$, or arbitrary continuous densities. Most relevant for speech processing are *Gaussian mixture models* which are defined as a weighted set of multivariate Gaussians:

$$p(x) = \sum_{k=1}^{\infty} c_k \mathcal{N}(x, \mu_k, C_k) \quad \approx \quad \sum_{k=1}^{M} c_k \mathcal{N}(x, \mu_k, C_k) \tag{3.26}$$

  To ensure a proper probability mass the components weights are non-negative and need to sum to one. The basic motivation behind such mixtures results from the *central limit theorem*, which states that any distribution can be described as mixture of an infinite set of weighted Gaussians, under the assumption that the distribution depends on a large set of independent factors. For practical applications the approximation error has to be minimized by using sufficiently large number M of component densities. In typical ASR implementations this number varies between several dozen up to several thousands.

Three basic problems of interest must be addressed when dealing with HMMs (cf. [Rab89, p. 270]): the alignment of an observation sequence against a given model, the computation of a matching score of this observation sequence for a given HMM, and finally the question of how to estimate the parameters of an HMM based on a set of observed feature sequences. These

```
Let $\alpha_t(i) = P(x_1, x_2, ...x_t, s_t = i)$
    1. Initialization
       $\alpha_1(i) = \pi_i b_i(x_1)$
    2. Recursion
       for all times $t$, $t = 1, ...T - 1$
       $\alpha_{i+1}(j) = \sum_i \{\alpha_t(i)\ a_{ij}\ b_j(x_{t+1})\}$
    3. Termination
       $P(\mathcal{D}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$
```

Figure 3.4: The forward algorithm, that computes the joint probability of an observation sequence $\mathcal{D}$. Basically, it accumulates partial path probabilities $\alpha_t(i)$ while traversing the observation sequence, which gives the joint likelihood for $t = T$.

problems are commonly referred to as *decoding*, *likelihood computation* and *parameter estimation* respectively.

1. **Evaluation Problem** Given a sequence of observations $\mathcal{D} = \{x_1, ..., x_T\}$, how to calculate the joint probability of all sequence elements $P(\mathcal{D}|\lambda)$? Concretely, this is the probability that the model λ has generated the observation sequence $\mathcal{D}$.

2. **Decoding Problem** What is the most likely state sequence $S = [s_1, ..., s_T]$ given a sequence $\mathcal{D} = \{x_1, ..., x_T\}$ of observations?

3. **Learning Problem** How to adjust the parameters of λ to maximize the joint probability $P(\mathcal{D}|\lambda) = \prod_{t=1}^{T} P(x_t|\lambda)$?

Fortunately, highly efficient and powerful solutions have been proposed for all three problems [Rab89]. These are the *Forward-Backward algorithm*, the *Viterbi algorithm* and the *Baum-Welch parameter estimation algorithm.* Their computation schemes are outlined in figures 3.4,3.5 and 3.6 respectively.

3.4.1 Parameter estimation

Beside Baum-Welch training depicted in figure 3.6, many derived parameter estimation techniques have been proposed. These commonly take other aspects like computation time or discriminative properties into account.

Baum-Welch training determines a probabilistic assignment of features to states, that allows a weighted re-estimation of the state OPDFs. A computationally much simpler, but in many cases equally well performing training scheme is to use a discrete feature assignment. For instance, the Viterbi algorithm allows to obtain such a discrete alignment of observations to states (cf. [Fin03] [Hua01]), and is thus the corresponding training scheme is referred to as *Viterbi training.*

The performance of estimated models strongly correlates with the quality of the initialization model. Because all EM schemes as BW-training iteratively optimize the model parameters, a poorly chosen initial model is unlikely to converge against the global maximum. Especially OPDFs are very sensitive to initialization. Because they have the largest impact on the performance of an HMM-based ASR systems [Rig98], they need to be initialized with special care. Most popular for

```
Let $\delta_t(i) = \max_{s_1,s_2,s_{t-1}} P(x_1, x_2, ...x_t, s_1, s_2, ...s_{t-1}, s_t = i|\lambda)$

1. Initialization
   $\delta_1(i) = \pi_i b_i(x_1)$                              $\phi_1(i) = 0$

2. Recursion
   for all times $t$, $t = 1, ... T - 1$
   $\delta_{i+1}(j) = \max_i \{\delta_t(i)\ a_{ij}\ b_j(x_{t+1})\}$         $\phi_{i+1}(j) = \arg\max_i \{\delta_t(i)\ a_{ij}\}$

3. Termination
   $P^*(\mathcal{D}|\lambda) = P(\mathcal{D}, s^*|\lambda) = \max_i (\alpha_i(i)$
   $s^* = \arg\max_j \delta_T(j)$

4. Backtracking
   for all times $t$, $t = T - 1, ...1$
   $s_t^* = \phi_{t+1}(S_{t+1}^*)$
```

Figure 3.5: The Viterbi algorithm, which reveals the best matching state sequence S^* given an observation sequence $\mathcal{D}$ with respect to an HMM λ. For an annotated description refer to [Fin03], or to [Jel97, p.45] for a more theoretical discussion. Similar to the forward algorithm, path probabilities are accumulated while iterating over the observation sequence. In contrast to the former, only the best path probabilities are kept (step 2), along with a backtracking pointer that subsequently allows to reveal the optimal path s^*. In each recursion step (2) all but the locally optimal partial path can be neglected as a result of the *optimization principle of Bellman* [The03].

initialization is k-Means as described in section 3.2 combined with Gaussian Splitting [San98].

3.4.2 Model adaption

A common problem in pattern classification are non-static variations of the input patterns. In case of speech these are caused by speaker changes, different accents, or varying background noise conditions. To cope with such changes, HMM adaption techniques have been proposed which aim to improve detection performance by altering the model parameters using small adaption samples gathered online.

A theoretically elegant way to adapt the parameters Λ of an HMM is to treat it as random variable, and to evolve it incrementally as new adaption data becomes available. This is referred to as *recursive Bayesian learning* [Dud00]. Let $X_1^n = x_1, x_2, \cdots, x_n$ a sample of size n. Than the posterior probability density $p(\Lambda|X_1^n)$ can be computed according to Bayes rule and by assuming all samples as independent by

$$p(\Lambda|X_1^n) = \frac{p(x_n|\Lambda)p(\Lambda|X_1^{n-1})}{\int p(x_n|\Lambda)p(\Lambda|X_1^n)d\Lambda} \tag{3.31}$$

Briefly, this makes a flat initial distribution $p(\Lambda)$ to converge against a Dirac delta distribution around the true value of Λ. However, this clearly only holds if there is only one Λ that causes $p(x|\Lambda)$ to fit the data. In such a case $p(x|\Lambda)$ is said to be *identifiable.* Because of the non-observability of the state sequence, there are serious computational difficulties to implement equation 3.31. Thus, approximations and assumptions are necessary to derive computationally feasible recursive Bayes solutions (cf. [Ma02] for a review).

Let

$$
\begin{aligned}
\gamma_t(i) &= P(S_t = i|\mathcal{D}, \lambda) = \frac{\alpha_t(i)\beta_t(i)}{P(\mathcal{D}|\lambda)} \\
\gamma_t(i,j) &= P(S_t = i, S_{t+1} = j|\mathcal{D}, \lambda) \\
&= \frac{\alpha_t(i)\ a_{ij}\ b_j(x_{t+1})\ \beta_{t+1}(j)}{P(\mathcal{D}|\lambda)} \\
\xi_t(j,k) &= P(S_t = j, M_t = k|\mathcal{D}, \lambda) \\
&= \frac{\sum\limits_{t=1}^{N} \alpha_t(i)\ a_{ij}\ c_{jk}\mathcal{N}_{jk}(x_t)\beta_t(j)}{P(\mathcal{D}|\lambda)}
\end{aligned}
$$

1. **Initialization**
   Choose a suitable initial model $\lambda = \{\pi, \mathbf{A}, \mathbf{B}\}$. This can be obtained for instance by using the unsupervised clustering techniques as described in sec. 3.1.

2. **Optimization**
   Refine the model estimate $\hat{\lambda} = \{\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}}\}$.

$$
\hat{a}_{ij} = \frac{\sum\limits_{t=1}^{T-1} \gamma_t(i,j)}{\sum\limits_{t=1}^{T-1} \gamma_t(i)} \qquad \hat{\pi}_i = \gamma_1(i) \tag{3.27}
$$

$$
\hat{c}_{jk} = \frac{\sum\limits_{t=1}^{T} \xi_t(j,k)}{\sum\limits_{t=1}^{T} \gamma_t(j)} \tag{3.28}
$$

$$
\hat{\mu}_{jk} = \frac{\sum\limits_{t=1}^{T} \xi_t(j,k)\ x_t}{\sum\limits_{t=1}^{T} \xi_t(j,k)} \tag{3.29}
$$

$$
\hat{C}_{jk} = \frac{\sum\limits_{t=1}^{T} \xi_t(j,k)\ x_t x_t^T}{\sum\limits_{t=1}^{T} \xi_t(j,k)} - \hat{\mu}_{jk} \cdot \hat{\mu}_{jk}^T \tag{3.30}
$$

3. **Termination**
   Terminate if $\hat{\lambda}$ obeys a user-defined threshold criterion or goodness of fit measure. Otherwise replace λ with $\hat{\lambda}$: $\lambda \leftarrow \hat{\lambda}$.

Figure 3.6: The Baum-Welch algorithm for parameter estimation of HMMs with mixtures of Gaussians as emission densities. In addition to the partial path scores $\alpha_t(i)$ introduced in fig. 3.4, $\beta_t(i)$ denotes the joint probability of a partial path that starts in state i at time t. The statistic $\xi(j,k)$ denotes the probability that the k-th mixture component (cf. eq. 3.26) of the emission density $b_j(x)$ associated to state j has generated the observation at time t.

Although the MAP-training proposed by [Gau94] is suitable for online-learning it is not suitable for one-shot learning. To overcome the problem of too little model changes caused by very few training samples (e.g. just one speech segment to generate a new word model) the classical MAP-training formulas need to be adapted to the given task (cf. [FF03]). The component density specific mixture coefficients are usually defined by

$$\alpha_i = \frac{n_i}{n_i + \tau}, \quad n_i = \sum_{t=1}^{T} \frac{w_i \quad p_i(x_t)}{\sum_{t=1}^{T} w_j \quad p_j(x_t)} \quad . \tag{3.32}$$

Thereby p_i denotes a state OPDF component density, w_i its weight and $x_t, t \in \{1 \cdots T\}$ the training sample of size T. The relevance parameter which regulates the ratio between old model parameters and model statistics is denoted with τ.

A *recursive maximum-likelihood estimation* (RMLE) method has been proposed by [Kri93] and was successfully applied to tasks like associative learning in an embodied cognition model (cf. [Squ05]). This method implements an iterative, stochastic gradient solution to find the maximum likelihood solution for a given sample. Formally it evolves the estimate about the model parameters Λ as follows

$$\Lambda_{n+1} = \prod_{G} \left(\Lambda_n + \epsilon_n \cdot \mathbf{S}(\bar{Y})_n ; \Lambda_n) \right) \tag{3.33}$$

where ϵ_n is a sequence of step sizes satisfying $\epsilon \geq 0$ and $\sum \epsilon_n < \infty$. In [Kri93] this method was proven to converge in case of continuous-density HMMs (and with minor modifications to the proof as well for discrete OPDFs).

Probably the most widely pursued approach for HMM adaption is *maximum-likelihood linear regression* (MLLR) [Leg95]. It basically clusters all HMM states unsupervised into "classes" and estimates a transformation matrix for each class. MLLR has been reported to improve speech recognition with a comparatively small amount of enrollment data [Pl01].

However, speaker adaption comes along with the drawback that it increases the number of false alarms. This is because it aims to increase the average model likelihood [Foo97, p. 214].

3.5 Statistical language modeling

As elaborated in section 2.2.1 statistical learning is considered to be one of the core principles of speech acquisition. It is assumed to take place on different levels of speech unit granularity, and is implemented in different manners as phonotactics or transition models.

Computational models that encode the statistics of time-series have been subject to research since many years. Most popular in the context of speech processing and discrete time-series analysis are *n-gram models*, that attempt to model the probability of a symbol given an already observed sequence of symbols of length $n - 1$.

An n-gram model corresponds to a n-th order Markov chain. It deals with a class of random processes that incorporate a limited amount of memory without actually being memoryless [Jel97]. The probability $P(\mathbf{w})$ of a symbol sequence $\mathbf{w} = w_1, w_2, \ldots\, w_T$ of length T is calculated by

$$P(\mathbf{w}) = \prod_{t=1}^{T} P(w_t | w_{t-n-1}, \ldots\, w_{t-1}) \tag{3.34}$$

Because each factor of this product has to be estimated or at least approximated in some way from a training sample, the size of n is limited in practical applications to $n \in 1, \ldots 5$. However, even for extremely constrained application scenarios, the number of possible n-gram tuples for which probabilities have to be estimated is huge. For instance with a small alphabet containing 25 symbols the number of tuples with length 4 is $25^4 \approx 4 \cdot 10^5$. For applications like unconstrained speech recognition the number of words is around 100k. Thus, even for commercial models that are trained with billion examples [Lam02b], only a small fraction of possible tuples can be actually observed in the training data.

Therefore, approximative schemes are required to provide fall-back values for tuples that have not been observed in the training data. The basic idea of those schemes is to diminish the probability of observed tuples to obtain an amount of freely distributable probability mass. This process is called *discounting*. Such schemes are still subject to research (cf. [Fin03] for a more elaborate review), because n-gram models are known to improve the performance of large vocabulary ASR systems by an order of magnitude [Hua01].

A popular n-gram smoothing approach, that has been incorporated in this work is referred to as *Katz smoothing* [Kat87]. It implements a *backing off* approach, that is to fall back to more general distributions when an n-gram context has not been observed in the training data. Specifically, the n-gram context becomes diminished until it has been observed, to provide at least a rough estimate about the n-gram probability. The above mentioned discounting mass is then distributed proportionally to this generalized model. A full mathematical description can be found in [Fin03].

3.6 Automatic speech recognition

Automatic speech recognition is one of the most successful pattern recognition technologies that has already reached consumer markets. This includes customer-support via telephone, dictation systems, or speech-control interfaces for devices from mobiles up to military airplanes.

The central problem of ASR is to determine the sequence of words $\hat{W} = w_1 w_2 \cdots w_K$ corresponding to a sequence of speech feature vectors $X = x_1 x_2 \cdots x_n$. Formally $\hat{W}$ can be obtained by computing the maximum *a-posteriori* probability $P(W|X)$ by

$$\hat{W} = \arg \max_{W \in \mathcal{P}(D^*)} P(W|X) = \arg \max_{W \in \mathcal{P}(D^*)} \frac{P(X|W)P(W)}{P(X)} \tag{3.35}$$

Thereby $\mathcal{P}(D^*)$ denotes the power set over all multi-sets of words in the dictionary D. Formally, all possible word sequences need to be evaluated in order to determine $\hat{W}$. Although $P(W)$ could be dropped from the actual computation because it does not depend on X, the maximization is computationally not feasible as it, because the number of possible word sequences is infinite.

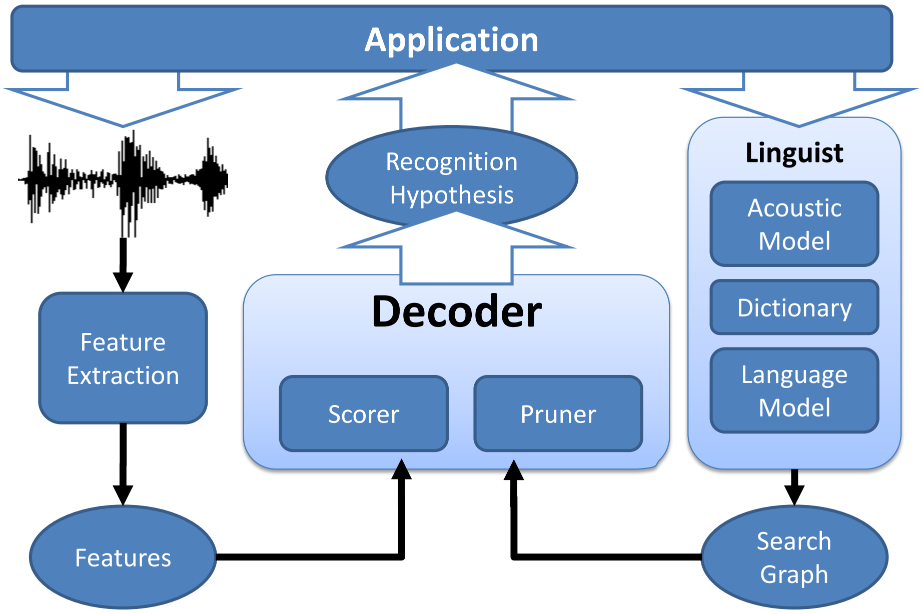


Figure 3.7: Core elements of an automatic speech recognition system. The feature extraction module converts the acoustic speech signal into an appropriate feature sequence. The linguist compiles a given grammar model into a search graph by converting each word into an appropriate HMM. The decoder aligns the feature vectors against the nodes of the search graph to reveal the best matching state sequence. By traversing this state sequence the best matching word sequence with respect to the given grammar model is obtained.

The practical challenge when building ASR systems is to design and estimate discriminative acoustic models $P(X|W)$ and language models $P(W)$ that constrain the search process to a computationally feasible size. The basic idea is to implement a parsimonious hypothesis search that neglects the overwhelming number of possible utterance candidates and examines only those word sequences as suggested by the acoustics [Jel97]. Most important are pruning strategies, that only keep a fraction of all possible paths in memory, while decoding an utterance as depicted in figure 3.5. This reduced set is referred to as *beam*, and the decoding approach itself as *beam search*. This reduction is as simple as powerful, and defines a core feature of almost all current ASR systems.

Even being a performance trick, such a pruning is to some extent even plausible from a biological point of view: It has been shown, that the brain ensures by means of lexical competition that only words that make up a consistent segmentation of the speech stream can be activated [Dav01], [Elm90], [Nor94].

Whereas HMMs provide the conceptual framework to model $P(X|W)$, they do not provide the means to build a functioning speech recognizer. It was subject to many years of intense research to build today's systems that allow to recognize almost unconstrained speech [Hua01]. Some aspects have to be taken into special consideration. First, because HMMs provide a means for discrete time series only, an appropriate abstraction level for acoustic speech needed to be found. Second, a sufficiently fine-grained model topology has to be chosen while keeping computational efforts feasible. Third, although solutions to the canonical problems as mentioned in 3.4 are available, they tend to be not practical when dealing with extremely large models as required for speech recognition. Common solutions to all three issues, are outlined in the following subsections.

The three most important topology parameters are the number of states per speech unit, the number of component densities, and the connectivity of the network. For the great majority of

ASR systems, those are chosen heuristically. This is because automatic model selection approaches like *Variational Bayesian Estimation* presented in [Wat04] are computationally costly and have not yet been reported to outperform heuristic approaches.

For a speech segmentation task on a Chinese speech corpus [Tao02] has found 6 to 7 states to be optimal for syllable modeling. However, their findings suggest that speech segmentation accuracy is rather insensitive to the actual number of HMM states. [Sar04] found 7 states HMMs with single Gaussians as emission densities to be optimal for syllable modeling. A slightly different setting has been implemented by Murthy in [Mur04] who employed 5 states with 3 component densities to model syllables. Generally, [Hua01] suggested that for each second of speech 15 - 25 states may wanted to be used. The single exception is silence which may be modeled using a simplified model topology.

3.6.1 Feature extraction

Speech is rarely matched directly against an existing set of speech unit models to determine the best matching speech unit sequence. Instead the speech signal becomes projected into an appropriate feature space in beforehand. This projection pursues two major goals. First, to make pattern matching more feasible the dimensionality of the input signal becomes greatly reduced. Second, it attempts to reveal speaker independent cues that are specific for the patterns to be detected, whereas speaker dependent portions of the signal are being suppressed.

Most popular in ASR are *Mel frequency cepstrum coefficients* (MFCCs) as depicted in figure 3.8. Basically they are obtained by low-pass filtering the short time log magnitude of the speech spectra, followed by a critical band analysis. Finally a discrete cosine transform (DCT) gives the cepstral features. MFCCs can be interpreted as weighting coefficients of sinusoidal basis functions that approximate the short time spectrum.

MFCCs are often preferred to other features like *perceptual predictive coding* coefficients. However, MFCCs suffer greatly in noisy conditions. By taking spectro-temporal aspects into account when extracting features, speech recognition performance has been reported to improve especially under noisy conditions [Dom09]. In addition, MFCCs are conceptually less suited to represent plosives because of the stationarity assumption of the FFT.

Temporal aspects seem be more natural to cope with syllabic framing of speech as investigated by Greenberg and Kingsbury in [Gre97]. They proposed *modulation spectrogram* features that are designed to reveal modulation frequencies between 0 and 8Hz with a peak sensitivity at 4Hz, which was found to match the average rate of syllables in continuous speech. The first step to compute modulation spectrogram features is to decompose a signal using a critical band FIR filter bank. Subsequently, low-pass filtering, FFT, averaged log-energy normalization, another FFT and magnitude calculation are applied to each channel. Final features are obtained using a bilinear interpolation between the different channels to give a image-like feature representation.

Comparable to region of interest detection in image processing that reduces an image to a patch of interest, speech is commonly framed into chunks prior to pattern matching. Most prominent are *voice activity detection* methods that parse the signal into speech and non-speech regions. In most cases such approaches rely on either energy-thresholds or generic speech/background noise models to classify the signal accordingly [Hua01, chp. 6].

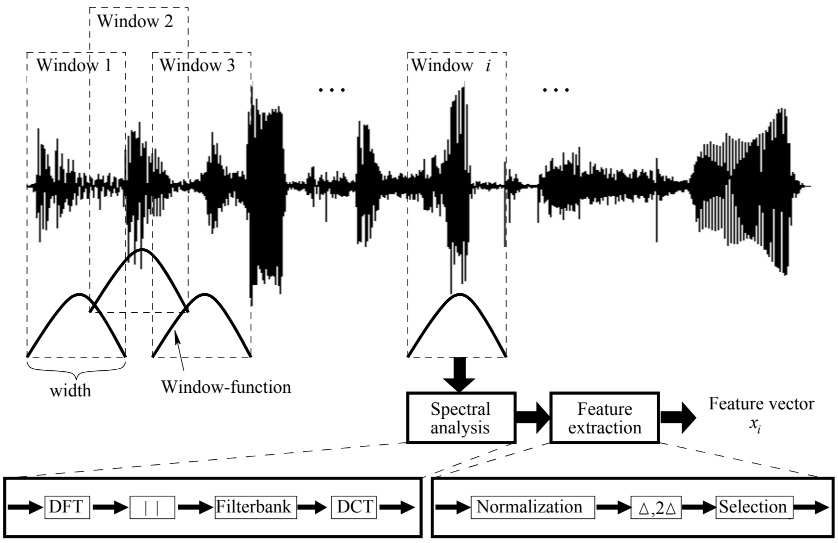


Figure 3.8: Processing steps when extracting MFCCs form speech. First, speech is framed into overlapping windows of approximately 25ms. Second, a frequency representation is obtained by transforming the framed speech signal into the spectral domain. This is followed by an auditory filter-bank analysis to reduce the dimensionality. Finally, the resulting feature vectors are normalized and become extended with time deviations of typically 1st or 2nd order.

3.6.2 Acoustic modeling

Speech is a time-evolving non-stationary signal. However, on a short time-scale it is assumed to be composed of quasi-stationary segments, that are modeled by state distributions associated to states of an HMM. As time evolves only in one direction, HMMs for ASR are designed with left-to-right topology [Hua01, sec. 8.2.4]. Because the number of feature vectors per utterance varies between different instances even of the same utterance, HMMs for ASR maintain self-transitions for each state. Furthermore, to cope with signal artifacts and noise, many HMM implementations include skip transitions that allow to draw paths that exclude some states.

Given a set of realizations of an utterance it is straightforward to apply Baum-Welch training introduced in figure 3.6 to estimate an HMM model for this utterance. However, scaling up this approach is not feasible because of the infinite number of possible utterances. Therefore, speech has to be modeled in chunks of sufficiently small size. These *speech unit* (SU) models can then be combined to give arbitrarily complex utterance models.

It seems to be computationally attractive to model speech in terms of phonemes. For instance around 40 phonemes are sufficient to transcribe almost every English word. Early ASR systems attempted to model each phoneme independently. This was considered to be most efficient because around 50 phonemes are sufficient to model all common languages.

However, due to co-articulation effects as delineated in section 2.2.4, these early ASR systems did not performed too well. Today's systems rely on HMMs that attempt to take those co-articulation effects into account. Because co-articulation is much less prominent across syllable boundaries, *context-dependent* HMMs are preferred for speech representation [Tol04]. The choice of context differs from tri-phones [Hua01], syllables [Wu97] [Shi97] [Wu98] [Nag03] up to complete sub-utterances. Such context dependent models are always trained with realizations from a similar context. However, due to limited speech corpora not all possible contexts are observable. Therefore appropriate heuristics and phonological knowledge are necessary to estimate acoustic speech

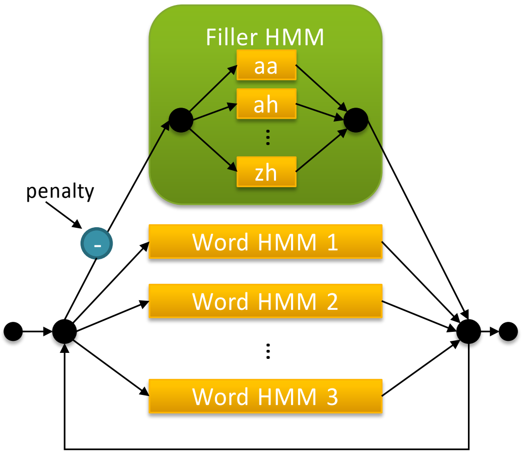


Figure 3.9: A filler-based keyword detection search space. Search paths are penalized when entering the filler-model by reducing their likelihood with a heuristically chosen penalty. Although not shown in the figure similar mechanisms are often used in ASR systems to penalize also word model transitions. This is necessary because acoustic likelihoods tend to outweigh language transition probabilities. Without such penalties short artifact words would appear frequently.

unit models. Commonly context-dependent models are clustered into groups according to phonologically motivated distance measures [Jel97, p209]. This allows to share training data between different models.

Some authors like [Pla92] or [Rus81] suggested to use *demi*-syllables instead of syllables for modeling of speech. Demi-syllables use shifted framing scheme, where boundaries are placed in nuclei locations. The main reason for such approaches is more the improved computational efficiency than a tighter connection to the speech processing performed by humans.

3.6.3 Keyword Spotting

As delineated in section 3.6.2 speech unit HMMs are compiled into a search lattice. Then, Viterbi-Decoding reveals the optimal alignment of states to a given sequence of feature frames. As long as the input speech fulfills the grammar specification this model has proven to result in utterance recognition results with high accuracy.

However, such systems cannot cope with out of vocabulary (OOV) words, which refer to parts of a speech signal that contains instances of words that are not part of the used dictionary. To overcome this limitation, special-purpose speech processing systems have been proposed (cf. [Tol04], [Mur04], [Tao02]). If the proportion of dictionary words to the overall amount of speech is low, such systems are referred to as *keyword spotters*. They are designed to match acoustic word models against a speech signal even if large chunks are composed from OOV word instances (cf. [Wei95]). If OOV words appear only occasionally, the task is to *skip* the corresponding segments but to ensure that subsequent speech parts are decoded correctly. Despite the different field of applications, such systems share the common idea to embed a dedicated *filler* model (aka. *background*-, *sub-word*- [Foo97] or *world*-model) between each dictionary word transition. Such filler-models are designed to cope all speech parts that are not described by any word model.

The design of the filler model is usually the tricky part. Different issues have to be taken into account. Such a model should be discriminative enough to give the highest path probabilities in the search graph on OOV-data, but should not interfere with (possibly dynamic) word dictionaries. The acoustic granularity of filler models spans from simple generic speech PDFs, over mono-phone-, multi-phone up to syllable- and word-models [Jun00] [Baz00] [Foo97, p. 212].

To avoid the usually tricky choice of a filler model Junkawitsch [Jun00] proposed a parallel decoder that continuously matches a set of keyword models to a feature stream in parallel. Models were chosen to be Hidden Markov models with *Bakis topology* containing 8 states. Each state modeled the feature space with a Gaussian mixture model comprising 4 component densities. Mel-frequency cepstral coefficients along with normalized energy extended with their first and second-order derivatives were used to give a 39-dimensional feature vector as input for the system. The output of each decoding processes was thresholded and fused to give a non-overlapping stream of segment hypotheses. However, this approach has not yet been shown to outperform filler-based systems.

As discussed elaborately in [Baz00], a serious problem when evaluating generic phone- or syllable-background models, is the high branching factor in the search beam. In contrast to dictionary-based search graphs only little language-model information can be incorporated because this contrasts to the desired generality of a filler-model. This means that more words sequences have to be taken into account when searching for the optimal word sequence according to eq. 3.35. To reduce the computational needs of KWS, Foote et al. [Foo97] proposed to match phone representations of the keywords onto phone-lattices to reveal the keywords segments. The lattice was yielded by normalizing the combined filler-keyword decoding results against accordingly aligned subword-models.

3.6.4 Confidence Measures

The result of HMM-based classification is a time-aligned sequence of models that maximizes the utterance likelihood. However, this does not imply that the recognition result is correct. Critical to reliable results are *confidence measures*, which approximate the probability that a recognition result is correct. This allows to reject unreliable recognition results, which is crucial for tasks like keyword spotting or OOV rejection. With respect to the subject of this thesis, such measures are mandatory to implement clustering algorithms in the speech segment space.

A conceptually plausible confidence measure is the posterior probability $P(W|X)$ as computed by equation 3.35. However, in actual speech recognition systems, only the nominator $P(X|W)P(W)$ is evaluated, because the denominator $P(X)$ is constant as it does not depend on W and can be accordingly neglected for maximization.

The most prominent approach to approximate $P(X)$ is to use a general purpose speech recognizer. Some works employ neural networks for this purpose [Ket06] [Ket07]. However, more rife are fully connected phoneme-networks, that are structurally equivalent to the filler models of section 3.6.3 [Baz00] [Baz01] [HT03, p. 214]: First, a Viterbi alignment on such a phone network is performed to yield $P(X)$. Second, a Viterbi-decoding has to be carried out on a word search-graph to obtain $P(X|W)P(W)$. Finally, keyword likelihoods are normalized with respect to the time-aligned filler likelihoods to give $P(X|W)$.

Related to this are approaches that rely on a set of "anti-word" models to approximate $P(X)$ by

$$P(X) = \sum_{W_A} = P(X|W_A)(W_A) \tag{3.36}$$

Anti-words can be also obtained from n-best lists [Jel97]. Having computed $P(W|X)$, a threshold criterion is usually applied to discard OOV-parts of the speech signal.

Other measures derived from the recognition hypothesis itself have been reported to further improve confidence estimation. These include word duration, language model scores, the number of phones in the recognized word sequence, or statistics derived from N-best-lists. For a more elaborate listing cf. [Hua01, Sec. 9.7.3]. The optimal weighting of these factors can be determined by linear discriminant analysis [Dud00, Chp. 5].

3.7 Neural networks for speech recognition

Artificial neural networks (ANN) map elements from an input space to an output space. As they allow to encode any kind of non-linear mapping, neural approaches have been also incorporated into ASR systems. Probably the most prominent model are *hybrid speech recognition* systems, for which state OPDFs are replaced by neural components [Rig98].

However, the actual use of neural approaches to decode a feature stream into a sequence of speech symbols remains an unsolved problem[1]. The great majority of neural approaches is not suitable for speech recognition due to the time-evolving character of spoken language. Neural networks have been shown to outperform HMM-based speech recognition in some cases for isolated word recognition tasks on small word dictionaries [Hua01, Sec. 9.8.1]. One suggested solution to overcome the limitations of strictly forward directed ANNs, is the use of *recurrent neural networks* that include connections between units to form time-delayed directed cycles [Bur94]. This allows such networks to keep an internal state, which makes it possible to match input sequences such as speech against an internal set of classes distributed over a larger period of time.

Although recurrent ANNs have been successfully applied in different domains [Elm90] [Dud00], applications to speech are rare because of their still very limited temporal memory [Wai89], [Noe91]. Unlike traditional RNNs, *Long Short-Term Memory* (LSTM) [Hoc00] attempt to avoid this problem, by circumventing vanishing gradients during training, and therefore can handle high as well as low frequency patterns. They were successfully shown to overcome some of the limitations of competing recurrent approaches also for speech applications [Ber04] [Fer02]. Finally, recurrent ANNs require different training and processing schemes compared to HMMs, and may behave even chaotically under certain circumstances [Bis08], which so far stunted their use in real-world speech processing systems.

[1] Formally, even if HMMs might seem to be not as biologically inspired as ANNs, they can be viewed as a particular kind of linear neural network (cf. [Bal94])

CHAPTER 4

Symbolic models for speech acquisition

Speech acquisition has fascinated researchers since ages. There is a large variety of models that enabled computational linguists to gain important insights into its underlying principles and processing schemes. Here we attempt to categorize such models into two different groups. First, we delineate models that treat speech acquisition on a purely symbolic level. Being computationally more feasible, such models have been shown to reproduce many findings from developmental psychology. Subsequently in chapter 5 we attempt to give an overview about models that directly deal with the acoustic speech signal. Such models are less elaborate with respect to developmental principles compared to symbolic approaches, which is due to the difficulties that come along when dealing with noisy input data.

4.1 Symbolic sub-syllable learning of speech structure

Many researchers share the idea, that phones define the basic unit of speech perception. Therefore, symbolic approaches for phone acquisition are hardly to find in the literature. This is because, without more fine-granular speech units, there are neither means nor any needs to learn phones in a data-driven way.

Nevertheless, there are a few works which attempt to capture the structure of phone sequences. Most prominent and recent are the works of Bazzi [Baz02], who proposed different approaches to learn phone n-grams. Even if his own motivation has been rather robust speech recognition and OOV-rejection than speech acquisition, his ideas are inspiring with respect to this thesis, because he especially investigated semi- as well as completely unsupervised approaches to capture structural constraints from phone sequences. The common theme of his work is the idea to replace a flat phone background model (as shown in fig. 6.7) by a more constrained one.

4.2 Symbolic syllable structure learning

In [Baz00] Bazzi investigated the use of bi-grams trained supervised to improve OOV-recognition performance. Together with Glass he described a methodology in [Baz01] to automatically derive a set of variable-length units to be used as OOV-model. The basic idea was to employ an agglomerative clustering approach that starts with individual phones that become iteratively merged to form larger units. The used distance measure was the mutual information between phone sequences. In each iteration step the pair of adjacent sequences/phones $\{s_1, s_2\}$ which maximizes $\mathcal{I}(s_1, s_2)$ was merged into one sequence $s' = s_1 s_2$. Bazzi argues, that such an approach gives syllable-like units at some point in time because intra-syllabic phone patterns tend to be more mutually dependent than inter-syllabic ones. However, because it is not clear how a possible stop criterion might look like, it seems hard to determine the iteration when syllabic units have been emerged. In the cited work, the number of iterations was chosen empirically.

One of the best performing approaches for syllable structure learning has been presented by Sharon Goldwater and Mark Johnson in [Gol05]. As usual for symbolic speech acquisition, they assume an innate representation of the input data on the next lower speech granularity level which are phones. Additionally they assume the sonority order of used phone symbols to be known a priori to the system. Their proposed model combines the sonority sequencing principle and the maximum onset principle as described in section 2.3 within a straightforward algebraic learning scheme to acquire syllable models. The resulting segment models becomes subsequently refined using an iterative EM training that attempts to maximize the predictive power of the model.

As baseline for evaluation they employ two different probabilistic context free grammars, namely a positional phone model and a generic phone order model. Although both are estimated on a annotated training sample, they are reported to perform less well in terms of F_1-measure than the refined algebraic model.

Contrary to bottom up syllable structure acquisition models are top-down approaches that split up words into syllables. The latter process is commonly referred to as *syllabification* and is trivial for languages with logographic writing systems like Chinese, where each sign represents a syllable.

For more fine-granular phonographic writing systems, this assumption does not hold. Syllabification is commonly achieved by applying a set of hand-crafted linguistic rules (cf. TSYLB2 for an example [Kah76]). This makes implementations specific to a particular language. Furthermore, rule based syllabification does not take speech rate into account, although it determines the syllable structure of language as described in section 2.3.

4.3 Symbolic word structure learning

Computational models for lexical learning face the same bootstrapping problem as infants do. Segmentation cues depend on the language being learned, but in order to develop cue extraction abilities a sufficiently large and stable set of seed items is required. However, some computational models have addressed this problem to determine which cues are relevant to solve which kind of segmentation problem.

The lexical learning model to be proposed in this thesis in section 6.5 is closest in terms of its computational approach to the work of Gambell and Yang [Gam05]. They proposed different models for lexical acquisition that integrate several cues. Speech was obtained by phonetically transcribing utterances of the Childes corpus [Mac95] using the CMU pronunciation dictionary.

Prior to processing, phonemes were grouped to syllables by applying the maximum onset principle as described in section 2.3.2. Their resulting evaluation set comprised around 260.000 syllables and was split into a test and a training corpus.

First, they evaluated how statistical modeling without any further cues performed in segmenting the syllable stream into words. Transition probabilities were estimated on the training corpus and word boundaries were inserted at each local minimum of the transition probability. The obtained word segmentation performance was reported to have a precision of 41.6% along with a recall of 23.3%. This means that 80% of all words were not extracted at all and 60% of all predicted words are no actual words. As pointed out by the authors, this low performance is caused by a structural deficit of statistical segmentation: Because word boundaries are postulated only at local minima, it is clear that a sequence of monosyllabic words cannot be segmented successfully. Previous experiments in [Saf96] circumvented this issue by using an artificial language consisting of 3-syllabic words. However, because around 90% of words in spoken English are mono-syllabic [Gre98], a local minima learner is unlikely to be segmentation strategy as used by infants. Gambell and Yang [Gam05] report a probability of 85% for two subsequent words being mono-syllabic.

By complementing statistical segmentation with stress information this drawback vanishes, and their proposed system achieves superior segmentation results compared to [Saf96]. They show that the performance of their local-minima learner can be dramatically improved to give a precision of 73.5% and a recall of 71.2%, when USC is taken into account as a supplementary segmentation cue. This shows that statistical segmentation cues benefit greatly when being complemented by what appear to be innate constraints on phonological structure. The basic processing scheme is as follows:

1. If two (strongly) stressed syllables are adjacent, a word boundary is postulated in between.
2. If there are one or more syllables between two strong ones then a word boundary is postulated where the pairwise transition probability reaches its local minima.

The evaluation of such a model was simplified in their setup, because the CHILDES corpus also includes stress information. However, because the identification of metrical patterns remains an unsolved problem, it is not clear how USC could be incorporated into a computational model that acquires a structural representation of acoustic speech as we aim to develop within this thesis.

The second model presented in [Gam05] implements a straight forward algebraic model, that exclusively relies on the principle of subtraction and the unique stress constraint (cf. sections 2.2.3 and 2.2.2). Slightly simplified the employed processing scheme was as follows:

1. If two stressed syllabled are adjacent, a word boundary is postulated in between
2. If known words enclose a sequence of weak syllables $W_{1:N}$, this sequence becomes learned as a new word.
3. Otherwise the word boundary lies somewhere in between W_1 and W_N and USC does not provide sufficiently rich information to segment $W_{1:N}$. Then a boundary could be guessed without adding any new items to the lexicon, or no boundary at all becomes is inserted.

This extends their first model but favors subtraction against statistical learning when segmenting weak syllables. This model also copes for monosyllabic words, and significantly outperforms the above mentioned statistical segmentation model. Furthermore, it provides a computational much less expensive framework compared to statistical learning (cf. section 3.5). Gambell and Yang reported a precision of 95.9% along with a recall of 93.4% ($F_{0.5} = 0.946$) which outperforms any previous study on the subject on data-driven word segmentation in a realistic setting. Given these results, algebraic or – as we denoted it in section 2.2.3 – subtraction learning seems to be a very powerful tool to reveal the word structure of speech. Furthermore, the authors argue that if *a learning strategy is simple, effective and linguistically and developmentally motivated, it is reasonable to expect that children do it too.* However, they do not abandon statistical transitions as such, but local minimum learning for lexical acquisition.

Higher order statistics based on transitional probabilities have been investigated within a word segmentation model proposed by Swingley [Swi05]. It condenses four different TP cues: TPs as defined in equation 2.1, TPs between adjacent syllable pairs and triplets, and the mutual information $\mathcal{I}(A, B)$ between adjacent syllable pairs. The latter is defined as $\log_2 \frac{p(AB)}{p(A)p(B)}$. This also incorporates the frequency of A in contrast to $TP(A \rightarrow B)$ [Saf96]. However, there is little evidence from developmental psychology that infants correlate high frequencies of syllables with subsequent word boundaries [Gam05].

To compute an actual segmentation the model incorporates a ranking scheme that maps all probabilities to percentiles and applies a set of decision rules to determine where to place word boundaries. A percentile filter threshold θ is optimized empirically. Due to the design of the classifier the model is not able to detect words with more than three syllables.

In contrast to most language modeling attempt the computation of TPs between syllable triplets requires a computational framing of 6 syllables. Whereas technically feasible, such an approach will require a vast amount of training data to accumulate reasonable statistics. Also LMs trained on large broadcast news corpora commonly assume a trigram context to be the upper limit for language model estimation. Possibly because of this inclusion of unreliably estimated higher order statistics, Swingley reports a quite low precision of $24 - 30\%$ along with a recall of around 25% when assessing segmentation performance.

As outlined in chapter 2, speech segmentation relies on a multitude of conceptually different cues. A common machine learning technique to integrate these cues is *co-training*. Its basic idea is to improve the performance of a supervised learning algorithm by incorporating large amounts of unlabeled data into the training process. Specifically, it implements a multi-classifier model initialized with a small amount of labeled data. For unlabeled examples, the assumption/hope is that easily detectable patterns for one classifier can be exploited to give further training samples for another. Co-training has been successfully applied to bootstrap part of speech (POS) taggers [Cla03] as well as for word sense disambiguation tasks [Mih04]. Voting schemes and confidence-filtering techniques used for other *Boosting*-techniques have been reported to further improve classification performance in such models [Dud00, chp. 9.5.2].

A popular variant of co-training is *self-training* (aka. *weakly*- or *semi-supervised learning*). In contrast to co-training only one classifier is employed to improve its own discriminative abilities [Mih04]. One important property of co-training techniques is, that they allow to bootstrap classifiers incrementally, and seem therefore suited to implement models for cognitive development *if* the annotated seeding sample can be gathered from by a developmentally plausible mechanism that relies on more basic processing mechanisms within a sensory processing hierarchy.

The principle of subtraction as detailed in section 2.2.3 has been implemented for phone sequences by de Marken [Mar95]. The central idea of this work was to cover an input utterance with a sequence of words from a previously acquired dictionary. Each word is represented as a sequence of phones. New words become created for uncovered portions of the utterance. Furthermore, this approach includes a regulative scheme that removes rarely used words from the dictionary.

A connectionist approach for lexical acquisition has been presented by Aslin et al. in [Asl98]. They employ a three-layer, feed-forward neural network to map a sliding window comprising 3 consecutive phonemes to an output unit that was activated at boundaries between utterances during training. When exposed to a test corpus the network was shown to predict not only utterance boundaries but also word boundaries *within* utterances.

Related the approach of Aslin is the work of Elman [Elm90] who proposed a recurrent network that was shown to predict phoneme symbols. Such a prediction approach differs from the above mentioned methods with respect to two aspects. First, it relies on phonemes for word segmentation and not on syllables. Second, it employs an inverted view on word segmentation that is related to the *entropy* based segmentation strategies synopsized in chapter 5.2. The evaluation data was a small artificial language presented as a segment at a time without word boundaries. Elman observed that the predictive power increased within words as more and more phonemes have been processed. Second, he reported that the predictive error increased sharply at the end of the word.

When applied to a large phonologically transcribed speech corpus the boundary prediction rate dropped to only 21% along with considerable amount of false alarms as reported in [Cai97]. Interestingly, [Cai97] reported that boundaries were rather placed between syllables than between words. This finiding supports a layered approach for speech acquisition as proposed in chapter 6.

Most works on lexical acquisition rely on single speech cues for segmentation. However, is seems reasonable to assume that infants make use of different cues in parallel to determine what and where word boundaries are. Computational evidence for this conjecture has been reported for instance in [Chr98]. By combining utterance boundary cues [Asl98], phoneme prediction networks [Elm90] and metrical stress [Cut87] to become the input for a recurrent neural net that was trained on a training corpus in beforehand, Christiansen and colleagues could show, that such a combined approach outperforms single cue (or pairs of cues) models. They could show that 74% of all word boundaries could be revealed correctly with the combined cue approach. However, two times more false alarms than actual word boundaries were predicted by their system.

Whereas neural approaches are well suited to match an encoded pattern against a feature context, they do not keep any stack of hypotheses as HMMs do, which would allow to recover from an initially incorrect classification and/or segmentation because of later evidence for a competing hypothesis. For instance [Nor94] reported problems using a recurrent neural net to recognize short words embedded at the start of longer words. Although HMMs are also likely to fail to recognize such short words immediately, the Viterbi decoding introduced in section 3.5 will usually recover to the correct the solution in order to find an (utterance-)optimal time alignment of features and states.

One idea to overcome this limitation of neural model is to extend the networks task that it must continue to activate identified words until the utterance offset [Dav01]. Davis' model was designed to circumvent supervised learning but focuses on utterance boundary information to build a discrete model for lexical acquisition. It implements an incremental lexical learning process based on a recurrent neural network approach. The model has been shown to successfully detect even onset-embedded words within a symbolic phoneme input stream. Furthermore the authors reported it to learn the relationship between speech and meaning without prior supervised training. For this purpose, the scene semantics were kept constant while processing each training utterance and the network task was to disambiguate the association structure of the vocabulary (cf. section 8.1). The network was shown to learn the structure of speech before it starts to map speech elements onto the correct lexical outputs.

However, the approach makes some assumptions that are unlikely to apply for infants. First, the used back-propagation training scheme does not account for the incremental way infants seem to acquire the lexical structure of language, but iterates several times over the same training set. Second, they used an artificial simplified test language that comprised only very few mono- and bisyllabic words with a CVC syllable structure. Third, a flat uni-gram model had been employed to generate the input utterances. Although this is clearly beneficial from a computational point of view because it makes all words to appear at the utterance boundaries, this assumption does not hold for natural language where certain words just appear within an utterance context.

Christiansen [Chr98] proposed a recurrent neural model for lexical learning. It first learns the statistics of the tutoring language by accumulating transition counts of phoneme series, and later exploits these constraints to segment speech into word segments. Christiansen evaluated the model on the CHILDES database [Mac95] which he supplemented with relative stress annotations taken from a lexicon, as well as special utterance boundary tags. Christiansen emphasized that the network was supposed to learn the regularities of phone-patterns at the utterance boundaries, and should be able to generalize this knowledge also to intra-utterance word boundaries [Chr98, p. 22]. In accordance with the above mentioned work of Gambell and Young, he supported the idea, that infants bootstrap lexical segmentation abilities by focusing on single word utterances in the speech stream.

CHAPTER 5

Acoustic speech modeling

In the last chapter we synopsized models that were driven by the motivation to capture structural constraints about speech unit boundaries using symbolic speech as input. However, because infants are faced with an acoustic speech signal instead, such models can only provide some insights into the processes that may underlay the infant's speech development. To overcome this limitation and to build a system that is actually able to learn the structure of speech in natural interaction with a human tutor, we have to bridge the gap between speech recognition technologies which most of us have already experienced in our daily live and aforementioned symbolic models.

This work would be rather short if speech recognition techniques could be straightforward applied to make symbolic approaches for speech structure acquisition to work also with raw acoustic speech as input. But as we will outline in this chapter there are major conceptual differences between common automatic speech recognition (ASR) systems and models for acoustic speech acquisition as researched in this thesis.

The main difference is the way the underlying speech representations are created. As outlined in section 3.6 ASR systems rely on annotated speech corpora to train the required acoustic models. Thereby, the quality of the alignment between speech units and the annotation is crucial for a high recognition performance. Manual annotation of speech recordings with phonemic labels and boundaries symbols almost always outperforms automatically obtained annotation with respect to the resulting recognition performance [Hua01].

Manual annotation of speech recordings is time-consuming and very expensive. This is due to the expertise required to produce speech annotations of high quality. Even experienced phoneticians require 20-30min to annotate one minute of speech. Another cost factor is the simple but widely accepted fact, that the performance of today's ASR systems increases with the amount of training data.

In contrast to ASR system developers, infants do not have access to annotated speech corpora. In chapter 8 we will investigate how contextual knowledge may provide the infant/system additional cues that help to ground its speech structure representation. But this is little compared to the rich annotation employed when training ASR models like tri-phone or syllable models. Due to this fact ASR model estimation techniques seem to be not suited at a first glance to build a model for infant-inspired speech acquisition.

Therefore it remains a scientifically and - maybe even more important - economically challenging task to develop acoustic model bootstrapping techniques that rely on as little as possible annotated speech. According to van Hemmert [vH91] such approaches can be broadly classified into *implicit* and *explicit* techniques.

Explicit (or *text-dependent*) methods time-align a speech signal against a known phonetic transcription. *Implicit* (also referred to as *text-independent*) techniques segment the speech signal into fragments, corresponding to phone-like (or syllable-like) units without any knowledge of a corresponding phonetic or textual transcription. Acoustic models resulting from explicit techniques tend to give the lowest error rates in ASR systems, since the number of detected segments equals the number of annotation symbols. This contrasts to implicit approaches were the number of predicted boundaries is not always correct.

To structure the wide variety of implicit segmentation approaches, we attempt to further categorize those into *model-based* and *direct* approaches. We understand the former as methods to bootstrap a representation in terms of fine-granular speech unit classes similar to phones or syllables. In contrast we consider *direct* approaches to implement signal processing methods that parse an acoustic speech signal into same-granular chunks like phones or syllables.

Direct methods are often not restricted to speech but can also be applied to other types of signal. This could be interpreted as an advantage. However, it is clear that machine learning schemes that include more knowledge about the pattern under question will tend to outperform less informed classifiers. Although direct methods involve certain amounts of domain-specific knowledge mainly imposed by the system designer, model-based methods encode a much richer description of the data-space under consideration. Therefore, model-based approaches tend to outperform direct ones in most scenarios.

The outline of this chapter is as follows. First, we summarize works that implement implicit model-based segmentation strategies. Subsequently we delineate direct approaches for speech segmentation. And finally, we summarize explicit acoustic model estimation techniques with a focus on works that implement adaption schemes to overcome the problem of few annotated training samples.

5.1 Implicit model-based speech clustering

As discussed in chapter 2, phones can be considered to be the lowest level of conscious speech perception. Although they clearly have time-dimension, they are often modeled as mixture distributions comprising between 1 and 8 normal component densities [Sha07]. The phone space is thereby spanned by the used speech features as delineated in section 3.6.1.

A very recent and powerful approach for unsupervised incremental phone learning and recognition has been presented in [Mar07]. It implements a network of states that is capable of unsupervised on-line adaptive learning while preserving previously acquired knowledge. Similar to ART networks [Gro88] it extends its representation autonomously by adding new states if the current input is considered to be sufficiently new[1]. State distributions are chosen to be unimodal Gaussians with fixed variance. The network connectivity is updated dynamically by using an aging scheme similar to growing neural gas as described in section 3.1.1: The age of the edge between the two most active states is set to 0 and the age of all edges becomes incremented by one. Edges that

[1] Even if the choice of appropriate system parameters is always a hard one, we were astonished by the *vigilance* threshold used in [Mar07] that was chosen without further elaboration to be $\ln \theta = -12 \ln(2\pi e)$

exceed an age-threshold as well unconnected edges become removed.

Mporas and colleagues presented a method in [Mpo08] to align speech waveforms to their corresponding phone sequences without exploiting any phone boundary information. Their basic idea was to instantiate a flat-initialized phone HMM for each training utterance using a shared phone inventory followed by an embedded iterative model re-estimation. These refined models are subsequently used to calculate a forced-alignment of the speech data, that defines the starting point for an isolated phone model training. The latter step is performed iteratively to further improve the quality of the resulting phone representation. However, the approach uses phone labels to construct the initial utterance HMMs, and is accordingly not suited to acquire phone models within a developmentally inspired architecture as targeted by this thesis.

Another approach has been proposed by Iwahashi in [Iwa03], [Iwa04] and [Iwa06]. His model is built around speech unit HMMs with three states and left-to-right transitions only. State OPDFs were chosen to be mixtures of Gaussians with eight component densities. A number of such unit models were embedded into a generic speech model HMM, in which transitions were allowed from final to initial states only. Using approximately one minute of input speech Baum-Welch training was applied to estimate the all model parameters. To estimate the required number of unit models Iwahashi proposed to use the Bayesian information criterion.

Qiao and colleagues presented a segmentation model in [Qia08] that assumes each phone to become generated by an independent source. They proposed a fast generic segmentation algorithm, that implements an agglomerative clustering scheme to merge adjacent frame-segments to larger chunks. Although very interesting and especially inspiring with respect to the used evaluation metrics (cf. section 7.1) their method assumes the number of phone segments per utterance to be known in advance. Clearly, this does not hold, when infants build up their speech representation.

In [Koh88] Kohonen presented a self-organizing map approach to learn a phonotopic map of unconstrained input speech. The trained network was shown to give reasonable trajectories when evaluated with test utterances.

5.2 Direct methods for speech segmentation

Most direct approaches involve two steps to segment a speech signal. First, it is converted into a highly sub-sampled *detection function* $A(t)$. This conversion is usually realized by digital signal processing techniques that attempt to emphasize changes in terms of harmonicity, spectral distribution [Ala99], phase deviation [Nag03], orthogonality [Foo01], entropy, pitch [Sat03], formant frequency [Tah01] or – most frequently – energy [Pfi96] [Jit98]. For instance, one popular [Jit98] realization of $A(t)$ is the *square energy* $E(n)$ of a speech frame n assuming a window size of W given a signal $S(t)$, that is computed by

$$E(n) = \sum_{i=1}^{W} = S(W \cdot n + i)^2 \tag{5.1}$$

Second, a peak picking method is applied to identify minima/maxima of $A(t)$. This is necessary to determine actual segment boundaries. Prior to this process, pre-processing techniques like low-pass filtering, median smoothing [Jit98] or normalization techniques are used to smooth the detection function $A(t)$ (cf. [Bel05, p. 1043]). Formally, a first order difference function $D(t)$ is derived from $A(t)$ by

$$D(t) = \frac{dA(t)}{dt} \tag{5.2}$$

Subsequently, normalization (e.g. with respect to the amplitude envelope of the signal) gives the *relative difference function*, that indicates the amount of change in relation to the signal level [Kla99].

Albeit many methods solely focus on temporal aspects, multi-band analysis has been reported to improve segmentation performance considerably (cf. [Bel05] for a review). For instance, a text-independent method for phoneme segmentation has been proposed in [Ave01]. The approach implements a preprocessing scheme along with a boundary detection mechanism. Preprocessing involves $20ms$ framing, band-pass-filtering, and equal loudness compression of the Fourier spectrum. Segment markers in different frequency channels are computed using a local maximum finder in the first time-derivative of the features. Finally, segment boundaries are chosen to be cluster centers of quasi-simultaneous changes in the different frequency channels. This is computed by accumulating all segmentation markers and detecting local maxima in the resulting function.

Clearly, such methods cannot be supposed to extract phonemes due to their rather linguistic than perceptual definition. However, some works on speech unit segmentation tend to ignore this difference and claim to segment linguistic units like phonemes or syllables without any plausible reasoning.

Pwint et al. proposes a maximum entropy segmentation approach to determine the number of segments in an utterance [Pwi05]. The segmentation markers are obtained using evolutionary optimization: A set of randomly initialized boundary markers is evolved to improve segmentation with respect to a target function that minimizes intra-segment and maximizes inter-segment homogeneity. Evolution is performed in terms of *crossover* and *mutation.* According to the authors *the proposed method detects consecutive digits as one segment only, when there is no inter-word silence between them* when being evaluated on the TDIGITS corpus [Leo93]. However, according to our findings in chapter 2, silence intervals between adjacent words are rarely present in spoken language.

Most direct segmentation approaches target segments of phone-length. Only few works attempt to detect larger units like syllables. The evaluation of such models is to some amount eased by selecting the evaluation scenario carefully. Particularly Asian languages with their clear and consistent syllable structure tend to be more easy to segment compared to European languages [Hsi99].

A direct syllable segmentation method has been proposed in [Nag03]. *Group-delay* features – computed as negative derivate of the Fourier transform – are extracted from speech followed by a local maxima detection. Evaluated on SWITCHBOARD this approach was shown to give boundary marker insertion and deletion errors of 5.25% and 7.1% respectively.

5.2.1 Model-based changed point detection

The main motivation to replace direct methods with model-based approaches follows a fundamental law of pattern recognition, that detection quality generally increases if more knowledge about the system under question is incorporated into the classifier. Therefore it is reasonable that forced alignment schemes [Rab89] for model estimation tend to give the best results. However, as outlined above, our focus is on bootstrapping representation in a solely data-driven developmentally plausible process. Thus, we emphasize on approaches here, that assume less and solely developmentally plausible innate knowledge, instead of linguistic information such as orthographic or phonetic transcriptions.

The least informed speech classifiers that nonetheless implement an actual model of the speech stream are probabilistically motivated 2-class models. Conceptually these assume the speech signal to become generated by a small set of generative probabilistic source models. For instance let us assume a vowel and a consonant model with the probability densities $s_V(t)$ and $s_C(t)$ respectively. Then a log-likelihood ratio can be defined by

$$s(V,C) = \log \frac{s_V(t)}{s_C(t)}. \tag{5.3}$$

The expectation of the observed log-likelihood depends on which model the signal is actually following. Given s_V as current model, the expectation is

$$\mathcal{E}_V[s(V,C)] = -\int p_V(t) s(V,C) = \mathcal{D}(p_V||p_C). \tag{5.4}$$

To detect a change in the signal, the integral is computed over a sliding short-time window. If the signal switches from V to C and vice versa, $s(V,C)$ will change its sign, from which a segmentation marker can be derived.

Another type of probabilistic segmentation models are *surprise functions* which are designed to highlight unexpected changes in the speech input with respect to an either local or global signal model. The amount of "surprise" is derived from likelihood measures or Bayesian model selection criteria [Abd03]. Hereby, the speech signal is rated in terms of a conditional probability which is conditioned by the already observed samples.

Beringer [Ber04] proposed a psycho-computational model for human phoneme acquisition. Especially, she suggests to use LSTMs (cf. section 3.7) to process an input signal comprising articulatory and prosodic features. As LSTMs were shown to give reasonable tessellations of arbitrary input spaces in other domains, they also seem to be suited to reveal the inherent structure of a semi-symbolic speech input. Beside the interesting idea, no data or results were reported in the referenced work.

A computational model that employs 5 different support vector machines [Dud00, chp. 5.11] to progressively segment continuous speech into vowels, consonants, fricatives, stops and silence has been proposed in [Jun03]. Each SVM computes a probability for the respective sound class. Thereby, sound-classes (e.g. speech sounds are either sonorant or not, non-sonorant sounds are either fricatives of stops) are presupposed to embed into a hierarchically ordered tree. This allows to compute conditional probabilities for each sound class, which directly provides the necessary mean to classify the input frames. The feature-space of each SVM was chosen to be a subset of zero crossing rates, frequency-band energies, pitch, and ratios between those.

A similarly cascaded approach has been pursued by Wang in [Wan03], where different feature sets were used to segment a speech signal into vowel, consonant and pause segments. Features were chosen independently for each classifier, and became selected heuristically from a wide range of speech properties like pitch, energy and distribution properties. The actual classification into utterance-segments was based on a set of hand-crafted rules along with a downstream AdaBoost [Bis06, chp. 6] to further improve classification results.

5.2.2 Syllable segmentation

Beside the few above mentioned methods, most syllable detection and segmentation approaches implement at least some kind of generic syllable model. Closest in terms of methodology and used representation to the model proposed by this thesis is the work of Murthy [Mur04]. He proposed a batch learning scheme for syllable-like units. It implements a grouping process of similar speech segments to define syllable HMMs, but lacks of the possibility to train models in a time-incremental manner.

A bracing new approach for syllable segmentation has been presented in [Hsi99]. First, speech input is labeled frame-wise by a hybrid neuro-fuzzy classifier to be either silence, consonant or vowel. Features were chosen to be zero crossing distances [Hua01] and an estimate of the first formant. The actual segmentation was performed on the symbolic label-sequence which exploits the fixed C-V structure of Chinese syllables. To solve the problem of vowel-vowel concatenations without any intervening consonants or silence, they propose a self-tuning back propagation neural network (STBNN) that requires energy and the time derivative of the envelope of the log spectrum as additional features. Evaluated on a Chinese speech corpus 93.1% of all frames became correctly classified. The STBNN could not be shown to significantly outperform a rule-based classifier that encoded a small set of heuristically found segment boundary patterns. According to the authors this was due to the limited amount of V-V concatenations in their evaluation corpus.

A neural approach investigated by Shastri et al. in [Sha99] employs a temporal flow model (TFM) to parse continuous speech into syllables. A TFM is a feed-forward network that allows also recurrent links, intended to smooth and differentiate the input signal. To overcome the problem of a too limited network-intrinsic temporal memory, several adjacent feature frames are grouped by a sliding window. A local dynamically adapted maximum-peak picking algorithm was applied to the low-pass filtered activity of the network to give syllable onset markers. To further improve performance, heuristics about *typical* (cf. section 5.3) syllable durations were used to prune unlikely markers. Modulation spectrogram features became employed as input (cf. section 3.6.1). Evaluated on a small set of 33 syllables a total onset accuracy of 84% (computed as sum of false-negative plus false-positive ratio) was obtained.

Because of the fixed input framing and the fixed time scale, the application of this model is limited to evenly sized syllables. However, a slightly modified more onset-centric approach could make this drawback negligible.

As discussed in section 2.3, syllable onsets tend to be more preserved compared to nuclei or coda (cf.). Thus, some systems have been proposed, that directly attempt to detect syllable onsets without performing an explicit syllable classification. For instance, multi-layer perceptrons have been reported to outperform direct segmentation approaches on this task by around 15% in accuracy [Mei99].

Furthermore, it has been suggested to make use of additional linguistic knowledge about the language to be segmented. For instance, by considering the demi-syllabic structure of Chinese along with language-dependent syllable cues like pitch contours, zero-crossing rates and energy-contours in an HMM-based segmentation system, it has been reported that segmentation accuracy significantly improves [Tao02].

A three layer perceptron for syllable onset detection has been described in [Shi97]. The MLP was trained supervised to distinguish between onset and non-onset features frames, which were chosen to be a hybrid set consisting of RASTA-PLP and spectral cues both sampled with $100Hz$. This was motivated by (assumed) synchronous rises in sub-band energy over adjacent sub-bands. To classify a frame while taking its context into account, four preceding and four subsequent frames were used as input for the network. Ground truth markers presented during training were broadened to five frames for trainability reasons. Evaluated on a continuous digits task, the system revealed an onset marker insertion error of 14.5% , which was reduced further to 6.378% by incorporating a minimal duration HMM-model. For Viterbi decoding on this model, heuristically chosen transition probabilities (cf. sec. 3.4) were complemented by local state likelihoods computed as negative logarithms of the MLP output activity corresponding to the respective state.

5.3 Word acquisition

As elucidated in chapter 2, words are unlikely to be learned as such without any precursory sub-unit acquisition. But this issue has been often ignored in the literature because of different scientific foci or lacking computational models for speech acquisition. Technically, word acquisition is often implemented by assuming predefined "innate" sub-word representations. This includes syllabic and/or phonemic representations trained on annotated databases [Roy00] [Bal03] as well as statistical models that encode phonotactic and syllabic word constraints. Moreover, by assuming the length of speech chunks to be a reliable cue to decide what a word is, word learning is assumed to be possible without bothering with linguistic principles at all. Although such an assumption is highly arguable, this idea is often pursued in word acquisition models due to the overwhelming complexity of developmentally plausible models for speech acquisition (cf. [Bra08]).

Probably the most well known application of word acquisition are dictation systems that require to enroll new words as names or subject specific terms. In most applications, users are required to pass predefined enrollment schemes, where the word to be learned has to be repeated several times without any speech context. This reduces the learning problem to a simple parameter estimation problem.

Another important application domain of word learning arises from the need to acquire new terms in interaction with a human robot. Thereby the focus is often rather on grounding and multi-modality than on developmentally plausible word acquisition [Roy00] [Bal03] [Iwa03].

Embodied word learning has been studied by Roy within the Cell framework presented in [Roy99]. It implements a multi-modal learning scheme where object labels and semantic categories are learned simultaneously. Cell implements an artful DTW-scheme that becomes applied to results of a phoneme recognizer in order to detect recurring phoneme sequences within a short time window. Such sequences are subsequently referred to as words, that additionally become associated to visual categories.

Cell lacks of a top-down feedback loop necessary to ensure a meaningful lexicon. Besides that, its speech processing back-end is an ANN-based phoneme recognizer, which was shown to be less powerful for speech recognition than context-dependent HMMs (cf. [Hua01]). Finally, the phoneme recognizer was trained on an annotated database. Therefore we consider the CELL framework to be a semi-supervised speech acquisition approach.

Recently, Cerisara proposed a model for unsupervised grounded word acquisition in [Cer08]. The main objective of this work was to build a phone-based semantic lexicon from a semantically enriched speech stream. Phoneme sequences obtained using a phoneme recognizer and were processed by a fuzzy string matching algorithm to reveal re-occurring sequences. The proposed algorithm does not exploit any prior knowledge, apart from a French phonetic recognizer constrained by a bi-gram model. The different processing stages of the system are:

1. Conversion of an raw audio stream into phoneme sequences
2. Automatic lexicon acquisition using fuzzy string matching methods to give morphemes
3. Topic clustering of the lexicon, by calculating the average distance of any pair of words in the speech stream. This allows to enrich word symbols with semantic tags.

Evaluated on a Broadcast news corpus against a manual word transcription baseline Cerisara reports a kappa value of $\kappa = 0.31$. Furthermore the approach allows to cluster audio-conversations into topic classes with$\kappa = 0.21$ which is at least far better than guessing. In the conclusion he outlines the idea of using acquired morphemes to improve recognition, which aims in the same direction as this thesis.

A similar approach has been proposed in [Bal03]. Speech is processed by a phone recognizer to give phoneme sequences, which are subsequently convolved to find recurring subsequences. Such are treated as words and are grounded within a multi-modal framework including action and visual object recognition.

The system of Iwahashi that has been already mentioned in section 5.1 also proposes an multi-modal framework for word-acquisition. The interactor is supposed to name object properties with isolated words. Based on a integrated audio-visual novelty function the system decides about to create a new word model, to update an existing one, or requests a confirmation if the novelty is less than a fixed threshold. New word models become created by simply concatenating the optimal sequence of phone models (cf. sec 5.1). Each acquired word is associated to either an action, an object or an object property.

Squire and Levinson proposed a system for grounded word acquisition in [Lev05] and [Squ05]. By following the idea that cognitive development can only occur through interaction with the physical world, they build an associative multi-modal model. This is realized by a cascaded set of fully-connected HMMs for each sensory input modality as delineated in figure 5.1. Each state corresponds to one category of the classifier. The decoded state sequences of the different classifiers are fused to give the input for an associative HMM. Although not limited to it, the authors evaluate their system using visual cues and speech as sensory modalities.

To make the approach work, the authors suggest three operational modes:

- **Auditory mode** where only auditory precepts are available and the system's task it to reveal the associated visual class.
- **Visual mode** which inverts the auditory mode by making only the visual input available to the system. The system's task is to reveal the corresponding acoustic class.

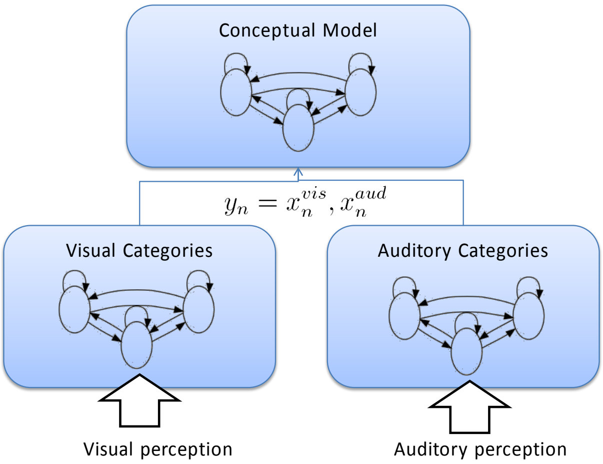


Figure 5.1: The cascaded HMM model for associative cognitive development as proposed by [Squ05]

- **Audio-visual mode** to bootstrap the associative layer using recursive maximum likelihood regression as described in [Kri93]

The model developed within this thesis shares the idea of using HMM-hierarchies of s to some extent. However, we consider a few aspects to be hardly evident for cognitive development or to be not completely clear from a technical point of view. First, the model does not suggest any means to actually learn the sensory classifiers, but uses predefined word- and visual-cue-models. Second, it is not clear to us how the used associative HMM differs from straight forward correlation learning. In our understanding the only difference is a penalty induced by the state transition probabilities in the associative HMMs. Functionally this should cause a smoothing of the decoded concept sequence. Third, the authors claim to use single states for each sensory class. These are expanded to independent HMMs for the auditory layer in order to make their system to recognize words. However, they do not propose any mechanism about how to learn these models in interaction.

Gold and Scassellati presented a framework for word acquisition in [Gol06]. It implements a minimal description length scheme to encode the phone structure of spoken language. It comes close to our framework presented in chapter 6 by bootstrapping a phone representation using a recursive hierarchical clustering scheme to find phone segments. Its outcome is a segmentation tree for each utterance. Leave-segments of different utterance segmentation trees are matched onto each other using a simple likelihood ratio test, which tags two segments as matching if the center frame of one segment produces a higher likelihood for a normal density estimated on the matching candidate than for a background model normal density estimated on the complete match-utterance.

For evaluation, complete utterances were processed that occasionally contained keywords to be acquired. A word classifier was trained by using a weak teaching signal that indicated which keywords were present in an utterance. This allowed to train an associative memory between teaching signal and co-occurring phone segments. Words were recognized when a referee's summed activity exceeded a threshold. Such activities are computed by weighting the segments of a test utterance with the associative weights as learned during training. The process relies on a high matching accuracy between phone segments found in the training data.

The authors report their system to find words reliably in continuous speech just after a few training segments. However, we doubt that the system scales up with respect to the number of words. This we think because of several reasons. First, the used associative model that encodes words, does not seem to take any sequence information into account (as opposed to HMM-decoding). Second, it lacks of an actual representation because each utterance is processed through the hierarchical clustering scheme followed by a matching against segments found earlier. This opposes findings from infant development, where it has been reported that infants get used to the phone inventory of their tutor's language (cf. section 2). Finally, they argue that some rules for word segmentation *follow naturally from the structure of the signal.* As an example, they claim that *k* or *t* are natural word delimiters. But this clearly does not hold. For instance this principle is violated by countless words in German language. If such rules would be really natural and therefore innate, infants of German-speaking parents would face are hard job to acquire their parental language.

5.3.1 Acoustic model bootstrapping

Because labeling of speech data is costly, researchers investigated methods to reduce the amount of annotated speech necessary to bootstrap ASR systems. The general idea of such approaches is to use a small sample for model initialization and to apply special EM-schemes to adopt the model structure iteratively.

Some authors have claimed to work in the direction of unsupervised acoustic model acquisition (AMA) [Kem99] [Lam02b] [Wes01]. However, most of these works present methods for acoustic model (AM) bootstrapping using a small set of annotated speech data: An initial AM is trained supervised with this annotated training sample and is employed to label a larger set of non-transcribed speech. These automatically labeled utterances are subsequently used to reestimate the model parameters. Sometimes this process is performed iteratively to further increase AM goodness. As stated in [Lam02b] *lightly supervised* AMA seems to be a more appropriate term for such approaches.

Similar *semi-supervised learning* schemes have been presented, where in a first step several classifiers are learned from a small set of labeled data. Subsequently, these classifiers are applied to unlabeled data, which provides additional training samples for additional classifiers to be created. A special case of co-training is *self-training* where automatically labeled utterances are used as new training samples for the labeling classifier [Lam02a] [Lam01]. Lamel et al. showed that the necessary amount of annotated speech could be reduced to as little as 10 minutes without major losses in ASR performance using such techniques.

CHAPTER 6

Model

How does language come to children? Scientists have investigated a wide variety of models to find an answer to this question. But so far any attempt to squeeze the complexity of speech acquisition into a computational model has been more brave than actually fulfilling its promise. As to our best knowledge no convincing computational model for lexical acquisition using acoustic speech as input has been proposed yet.

Discrete models as discussed in 4 revealed how speech acquisition principles can be mapped to powerful computational processing schemes. Such models are bootstrapped in most cases purely perceptually driven, and have been reported to give a high accuracy in segmentation, classification and clustering tasks. However, discrete approaches fail to provide a model for developmentally plausible speech acquisition, because they discard the nature of speech: In contrast to discrete symbol sequences, speech is an inherently noisy time-continuous function. This makes it hard to think up a possible transition from a symbolic to the continuous domain. Hence, symbolic models provide rather a source of inspiration when modeling speech acquisition, than an actual template to be slightly modified to process acoustic speech instead of symbol sequences.

As delineated in chapter 5, the large majority of works on acoustic speech acquisition focuses on isolated word learning. The majority of models neglects the underlying bootstrapping processes which acquire phones and syllables prior to words. Thus, existing acoustic speech acquisition attempts lack of developmental plausibility as more fine-granular representations seem to be mandatory for infants to uncover the word structure of their tutoring language (cf. sec. 2.2). Furthermore, there are only few reported models that account for the incremental manner in which infants learn language.

Even the most powerful acoustic bootstrapping approaches as described in chapter 5 have been reported to detect less than three quarters of words boundaries but a large amount of false alarms. Consequently, such systems fail to segment half of the words in continuous speech. Clearly, speech segmentation is conceptually challenging and computationally demanding given the complexity of the learning problem comprising a highly dimensional input space, co-articulation of adjacent speech units, various kinds of noise, speaker-dependence, and the huge number of model parameters to be estimated. Thus, it seams reasonable that no model has yet mastered to provide a plausible model for speech acquisition. Another explanation may be, that previous models neglected processing principles that are believed to facilitate speech acquisition in infants.

It is the aim of this thesis to fuse the best ideas of symbolic and acoustic speech acquisition into a common framework. The outline of this chapter is as follows. We first emphasize constraints and requirements for a developmentally plausible model of speech structure learning. There we discuss possible representations, plasticity issues, and possible schemes to link different subprocesses and representations. After a system overview we then describe in detail our model to bootstrap words, syllables and phones within a coupled hierarchy of incrementally bootstrapped HMM-layers. Finally, we try to embed the model into the scientific context and highlight differences as well as similarities to existing works as outlined in chapter 4 and 5.

6.1 Computational requirements and constraints

Any model has to focus on a subset of all possible aspects of the system under consideration. Therefore it is at first necessary to clarify what aspects of speech acquisition our model is aiming to reflect. Fortunately for us, these facets arise naturally because of our focus on *developmental plausibility.* This imposes severe constraints on computational principles, processing schemes and possible kinds of representation. We consider the following aspects to be especially important:

- **Type of input speech** How to represent and to encode speech within the model?
- **Speech representation** What are suitable computational representations for the different perceptual speech units?
- **Processing principles** What are the relevant computational mechanisms that enable infants to bootstrap speech abilities?
- **Order of bootstrapping** What is a plausible temporal order to bootstrap the different elements of the speech representation?
- **Coupling of sub-representations** What are the dynamics and implications when coupling different speech representations into an incrementally bootstrapped framework?

Clearly, this list is not complete as it disregards other important issues like multi-modality or necessary links to speech production. Both issues have deliberatively decoupled from this chapter to ease design and to increase readability. Our first attempts to tackle them are summarized in chapter 8.

6.1.1 Type of input speech

The input of our system needs to be the same as what infants perceive. This renders symbol sequences to be unsuitable because those usyually neglect fine-granular speech-unit bootstrapping processes. Hence, **acoustic speech** should be the input to our system. However, as physical mechanisms of sound transmission and reception are antecedent processes not related solely to speech, we consider speech as a feature representation that preserves the spectral properties of the speech signal.

This is supported from what is known about human sound perception (cf. [Dom09] for a review). Because of their proven performance in ASR systems, we consider Mel-frequency cepstrum coefficients (cf. sec. 3.6.1) to be a suitable encoding of the speech spectrum. However, as our framework is not supposed to impose any constraints on the speech representation, MFCCs may be replaced by more noise-robust and perceptually better motivated features as those become available (cf. chapter 8).

6.1.2 Speech representation

Probably the most important design decision when developing a computational model for acoustic speech acquisition is the choice of an appropriate speech unit representation. Previous works on speech acquisition employed often either neural networks or HMMs, whereby the former relied mostly on symbolic speech as input and had a stronger developmental focus (cf. section 4.3). In contrast – as elucidated above – HMM-approaches tended to be less developmentally plausible but seem to be better suited to model acoustic speech.

Because neural approaches have been successfully applied to build models for symbolic speech structure acquisition, it seems reasonable – at a first glance – to apply similar models to acoustic speech. This has been suggested for instance by [Fer02] who evaluated a recurrent neural scheme for digit recognition. However, despite the potential of such models we favor a Hidden Markov based speech representation because of several reasons.

1. **Scalability** Languages do not comprise tens or hundreds of words. [Sim89] includes around 600.000 word definitions. Clearly, the number of words in English is more a matter of definition than of calculation. More realistic estimates can be obtained from systems that aim to annotate raw broadcast news data. Such systems setup on dictionaries containing around 100k entries [Sch05]. Because news broadcastings can be assumed to be understood by most people, 100k marks at least a lower boundary for the size of the mental lexicon of adults. Thus, as present computational neural architectures have not yet been reported to scale up accordingly, HMMs seem to be favorable.

2. **Computability** Hidden Markov Models alone would not have become the predominant approach for speech recognition. Their real value comes from the clear and efficient computational framework that has been developed around them over the last two decades. This especially includes decoders, language model integration and training schemes (cf. [Rab89] or [Hua01] for a review).

3. **Time-series** Except from [Fer02] and a few works on phone recognition [Roy00] [Bal03], recurrent neural approaches have not yet been reported to succeed in real word speech recognition tasks. In our opinion this is mainly due to conceptual insufficiencies of RNNs: Although those are able to match time-varying inputs to previously learned patterns, they perform quite poor as the time duration of patterns exceeds more than a few processing frames (cf. sec. 3.7). However, as speech is commonly encoded in a feature space with a time-resolution of around 100 Hz, words comprise in most cases more than 100 frames. This is a magnitude more than today's recurrent neural network approaches can handle.

4. **Incremental learning** Because of search tree structures compiled by HMM-based speech decoders, it is straightforward to incorporate new models even during the decoding process. In contrast, neural approaches use a distributed representation which renders it difficult to become extended. So even if most neural learning mechanisms reflect with greater detail the way the human brain works, they rely on batch training schemes that do not support model adaption or incremental addition of new perceptual categories.

   Furthermore, according to Davis [Dav01] most current recurrent neural approaches require supervised training. As argued above, this clearly contrasts to how infants learn structural language constraints solely from perception and interaction with their parents. Hence, supervised learning seems to play a minor role while acquiring language abilities, which makes neural approaches less suited to model acoustic speech structure acquisition.

As any computational model, HMMs suffer from some structural inadequacies that make their use arguable. These include especially the first order assumption, exponential decaying state occupancies instead of a proper time representation, or the independence assumption [Bil04]. But most of these problems have been tackled in the past (mostly to the price of enormously high computational costs [Hua01, sec. 8.5]) without any significant improvements of speech recognition performance. Thus HMMs remain the first choice when building large vocabulary continuous speech processing systems [Jel97].

The second major issue of speech modeling is the question of *granularity*. According to [Dud00, Chap. 9.4], the granularity of speech units realized in ASR systems needs to be a trade-off between the following requirements:

- Units should be *accurate*, to catch the full structure of possible acoustic realizations.
- Units should be *trainable*. For each unit instance enough training data should be available. The less fine-granular a unit is, the more units are necessary to model the speech unit space.
- Units should be *generalizable*, so that learned unit models also match to new instances of the same speech element.

Probably the most popular unit used in ASR systems is the *phoneme* as discussed in section 2.4, that refers to a minimal sound sequence which turns one word of a language into another word. A complete sub-branch of linguistics named *phonology* deals with the issue to study language-specific patterns of sound and gesture and to finally select a set of phonemes for a particular language. To improve performance of ASR systems, derived units as *n-phones* are popular, which take co-articulatory effects between adjacent phonemes into account.

Whereas the above mentioned criteria for unit-selection are mainly motivated by practical and performance reasons, our focus in this thesis makes it necessary to add another mandatory condition for the units to be modeled.

- Units should be ***developmentally plausible***.

This implies some severe constraints on possible units. Especially it renders phonemes to be unlikely as perceptually units of speech perception. This is because infants seem to acquire the structure of speech in terms of phonological constraints and syllabic units prior to words as discussed in chapter 2. There, findings from developmental psychology indicated speech perception to be organized in terms of **phones**, **syllables** and **words**.

6.1.3 Order of bootstrapping

As discussed in chapter 2, words are likely to become acquired in a bottom up fashion starting from fine granular units that are linked up in some way to form more complex perceptual structures that finally cohere into words. For a computational model it is therefore reasonable to assume refinements in fine-granular representations to trigger higher level learning processes. This inspires the system architecture of this thesis to model speech structure acquisition as a **cascaded set of coupled bottom-up bootstrapping processes**.

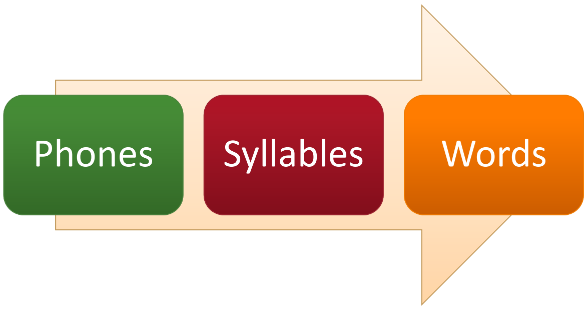


Figure 6.1: A developmentally plausible order of bootstrapping to acquire perceptual speech units of different granularity. As more complex units rely on linguistic knowledge not innately available, structural constraints on their constituents need to be develop prior to bootstrapping of higher levels of representation.

Phones can be considered as the lowest granularity level of conscious speech perception. They are rarely observed isolated even in child-directed speech. The number of phones is several magnitudes smaller compared to the number of syllables or words. Furthermore, phones are assumed to be learned without any non-innate linguistic knowledge. Thus, it is reasonable to assume that infants rely on unsupervised clustering techniques to learn a phone representation of their parent's language. This idea has been supported by most models of phone representation learning as delineated in section 5.

The natural extension to phones is a phonotactic learner that aims to capture structural constraints on what makes well-formed syllables in the tutoring language. As soon as such a set of structural constraints converges and thus allows to predict syllable boundaries, the actual learning of syllables can be initiated.

According to findings of developmental psychology summarized in section 2.2, word acquisition seems to be grounded in syllable perception. Hence, the learning of syllables has to precede word acquisition at least partially. However, as some syllables may appear only rarely and the number of syllables may be extremely huge, word and syllable learning processes are likely to co-occur to some extent.

Summarizing all three aspects gives the order of bootstrapping for our system that is outlined in figure 6.1. Phones learned prior to syllables. Syllable learning precedes word acquisition at least partially, because syllabic constituents that cohere into words need to be acquired prior to those.

6.1.4 Processing principles

A developmentally plausible model for speech acquisition should allow to resolve temporary ambiguities ensued by onset-embedded words (e.g. `cap` in `captain`). According to findings of Gambell [Gam05] and Davis [Dav01] such capabilities seem to be to facilitated by the learning ofstructural constrains about syllable (and correspondingly word) boundaries from the utterance boundaries. This is motivated because utterance boundaries are far easier to detect compared to intra-utterance word boundaries.

As outlined in chapter 2, speech acquisition is likely to rely on a multiplicity of mechanisms. Thus, models that integrate several cues seem to be better suited to model infant speech development. Although not commonly realized as supervision, the parameterization of such bootstrapping processes and their data-flow schemes are based on expertise, and could be regarded as kind of

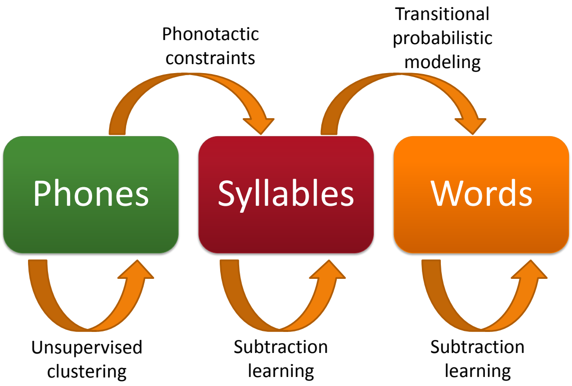


Figure 6.2: Developmentally plausible processing principles for speech acquisition implemented in our model. The dependencies between the different speech representation layers further motivate the bottom-up manner of bootstrapping as discussed in section 6.1.3

supervision. However, as without any assumption no model at all could be realized, so we consider mechanisms and their parameterization to be innate, without diminishing the notion of unsupervised developmentally plausible learning.

To acquire a phone representation, any unsupervised clustering approach seems to be suited, that is able to categorize high dimensional feature distributions. As we agree to the idea of similar works to regard phones as feature clusters with only a tiny time-dimension, inter-frame dependencies have to be taken into account only by little extent when estimating phone models. The only result that might influence the choice of a particular phone clustering method is the finding from developmental psychology, that infants narrow their phone perception to become more specific for the language being learned (cf. chapter 2).

As phonotactics seem to play the dominant role when acquiring syllabic constraints, our model should reflect this by implementing a phonotactic learning and parsing approach. Formally, phonotactic constraints allow to setup a syllabic parser. Following the ideas of section 2, this parser provides training segments for the syllable learning process, which needs to be complemented by a subtraction module, and a statistical learning model that allows to impose constraints on syllable transitions. Conceptually the word layer would be very similar to such a syllable layer, as it also constitutes from a unit detector, a statistical learning module and a lexicon learner that is based on subtraction learning.

Subtraction learning

When starting our research about a unsupervised acoustic speech structure acquisition, we relied on mono-syllabic utterances to facilitate the learning of new syllable models [Bra08]. This step was necessary for us to obtain deeper understanding of the underlying bootstrapping processes. A natural – and necessary – extension is to use detected segments as feedback signal. This can be realized by applying the principle of subtraction as described in section 2.2.3: Syllable detection results allow to derive residual segments which give additional training segments. By doing so, the model is able to profit from segmentation results obtained even under the assumption of incomplete speech unit representations.

For instance, given an already acquired model of the syllable `[si]` and a sequence of syllables `[a] [si] [mo]` as speech input, the syllable spotter will be able to detect `[si]` within this sequence. This allows to "subtract" the spotted syllable segment from the framing voice activity segment, which gives two residual segments `[a]` and `[mo]`. By incorporating those into the training process, the system's representation will fast converge against the underlying generative speech structure model of the tutor.

More technically, the idea is to apply the principle of subtraction to the results of the syllable detector which parses any speech utterance into detected syllables and background residual segments in case that no acquired syllable matches to a particular speech snippet. Such an approach is not restricted to syllable acquisition but is also applicable to lexical learning. The only difference is the level of speech granularity on which residuals need to be determined. In case of syllables these are sequences of phones. Accordingly residual segments on the word level will be composed by sequences of syllables.

Statistical learning

While mastering the task of speech segmentation, infants have been shown to learn the statistics of speech unit transitions [Asl98]. According to the findings discussed in section 2.2.1, these statistics seem to play an important role on *all* levels of speech granularity. As we aim to build a model for early infant speech development, we take such findings into account by accumulating transition statistics for phones, syllables and words alike.

As discussed above, speech development has to be modeled as an incremental process. Thus, the learning of speech unit transition statistics differs considerable from the offline estimation of language models as integrated into most ASR systems. Nevertheless, we can profit from the immense body of work in this field. Especially n-gram models summarized in section 3.5 seem to be a valuable and valid tool for our purpose, as they allow to represent contexts of arbitrary size and can be refined incrementally by simply updating unit tuple frequency counts. However, as outlined in [Gam05], such an update policy may comes along with a high computational cost due to the definition of transitional probabilities: If a learner observes a syllable A she must adapt the values of all $TP(A \rightarrow *)$ because of the absolute frequency of A in the denominator in the definition of TPs (cf. eq. 2.1). Thus, a direct implementation of statistical learning would require to update around 100k TPs for *every* processed syllable in a realistic speech acquisition setting.

6.1.5 Coupling of speech unit representations

A central question when building a speech acquisition model comprising speech unit representations on different scales is whether those should be organized hierarchically. At first glance it seems quite intriguing to model speech units as concatenations of more fine-granular entities. The problem is visualized in figure 6.3: words may be composed from syllables, which themselves may constitute by the concatenation of phone sequences. Using HMMs as computational model, such a concatenated approach would trivial from an implementation point of view. However, it fails because of two reasons.

First, a concatenation of basic building blocks like phones has been discarded by ASR research because of co-articulation effects previously discussed in section 3.6.2. Such a concatenation has been reported to result in a diminished performance compared to less granular units like words,

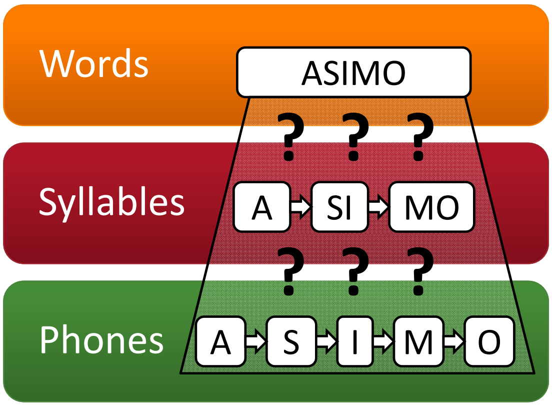


Figure 6.3: Decomposition of the word *asimo* to illustrate a possible hierarchical organization of speech perception. As indicated by the question marks, it is not clear from the findings of developmental psychology, whether perceptual units of higher levels are concatenations of more fine-granular units. In contrast, all speech units could also be represented as distinct entities without any hierarchical structure.

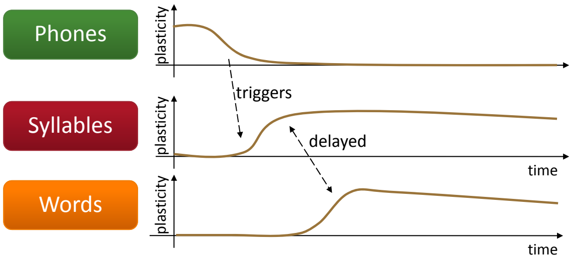


Figure 6.4: The temporal change of plasticity in the different sub-representations. As the number of phones in all languages is sufficiently small, phones can be assumed to be learned prior to words and syllables and to be kept fixed consequently. As phone model plasticity decreases, syllable learning becomes initiated. Because of the large number of syllables in most languages, the syllable representation plasticity needs to be maintained over a longer timespan and to decreases only gradually. Words are built from syllables. Thus, their acquisition is delayed until some utterances can be completely syllabified with the emerging syllable representation.

syllables or – most popular – tri-phones.

Second, the dynamics within each of the sub-representations need to be taken into account. If units are kept fully dynamic, dependent higher-level units may change their discriminative function over time. But as this would change associative mappings in an embodied context, a strictly hierarchical implementation seems hardly feasible using HMMs as speech unit representation.

To overcome these problems it is necessary to find a trade-off that allows to balance model stability against plasticity and diminishes co-articulation effects. Two different types of representation plasticity have to be considered. First, plasticity in terms of adaptiveness: Speech unit models have to be adaptable to new environment conditions like background noise or speaker changes. Second, it is necessary to maintain the extensibility of an existing representation with new speech units.

As the set of phones in most languages is comparable small compared to syllables or words, phones can be modeled using a straight-forward clustering mechanism prior to words and syllables. By disregarding changing speech reception because of aging effects, phone models can be kept constant after having been estimated. Such an approach can be regarded as a narrowing phone perception that becomes constrained to the phones used in the tutoring language.

To encompass co-articulatory effects, syllables need to be modeled as distinct entities. As initialization is crucial for HMMs syllables and phones are linked weakly by using concatenated phone models as initial syllable models (cf. section 6.4.3 for technical details).

Although it would have been possible to directly adopt the approaches of [Mar95] [Foo97] or [Roy00] to model words as phone sequences, we favor syllables as building blocks for three reasons. First, as detailed in chapter 2 most cues for word segmentation rely on syllables rather than phones. Second, inter-syllable co-articulation effects are far less prominent compared to those between subsequent phones. And as words are composed by syllables, word co-articulation can be considered to be affected by only few co-articulatory effects. As a consequence, *words are modeled as discrete sequences of syllable symbols* within our framework. Beside its developmental plausibility, this considerably reduces the computational and methodological complexity of the architecture without any major conceptual shortcoming.

This gives the acquisition scheme as delineated in figure 6.4. It outlines plasticity as a functions of time for all speech unit representations. It is designed to minimize the mentioned plasticity issues while implementing a partially hierarchical speech representation scheme. Phone plasticity is diminished to zero as soon a stable phone representation has been acquired. Subsequently, syllable learning becomes initiated. As words are built from already acquired syllable models, lexical learning is delayed against syllable learning. Another alternative would have been to delay word acquisition completely until the syllable representation has been converged. However, we favor the depicted approach here to avoid an unlikely long delay of the lexical acquisition process induced by the large number of syllables in most languages.

6.2 System overview

The proposed system architecture is shown in figure 6.5. Its design is solely driven by the constraints and requirements as discussed in the last section. Three interconnected layers are employed to learn the phone, the syllable and finally the word structure of an arbitrary input language. All three layers share a common structure: Each comprises a pool of HMM speech unit models, a speech unit detector and a statistical speech unit grammar. Initially all representations are empty. Processing and learning are organized in a bottom-up manner. The learning of phones and phonotactics completely priors syllable and word acquisition which allows to neglect stability and plasticity issues as discussed in section 6.1.5. In contrast syllables and words are acquired incrementally in parallel.

As shown in figure 6.5 the acoustic speech input is framed by a voice activity detector as described in [Wal04] into segments. These contain utterances of different complexity starting from isolated mono-syllabic words up to sets of utterances comprising many poly-syllabic words. Speech segments become converted to sequences of Mel-frequency cepstrum coefficient vectors including

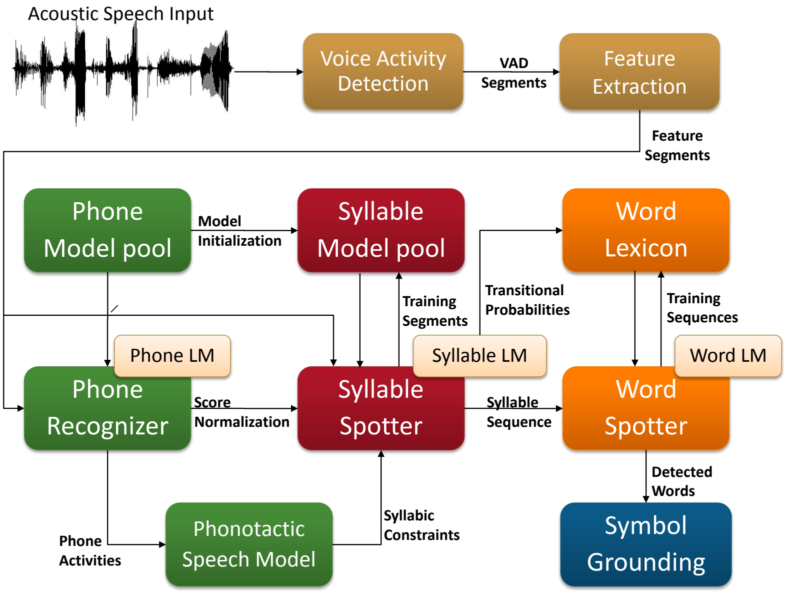


Figure 6.5: The proposed three-layered architecture for speech acquisition. As indicated by the visualization, all layers share a similar structure consisting of a pool of speech unit-models, a statistical grammar (LM), and a recognizer which detects learned units in the incoming speech stream.

energy, and their first and second time derivatives (cf. section 3.6.1).

Feature segments are processed by the phone and the syllable subsystem. Both aim on bootstrapping an appropriate acoustic unit representation. The phone layer integrates a phonotactically constrained Viterbi-decoder (cf. [Baz00]) that converts feature segments into phone symbol sequences. Such phone sequences define the input for a phonotactic learning module, which provides supplementary segmentation cues to the syllable layer. Training segments for syllable bootstrapping are obtained by means of phonotactically constrained subtraction learning. These segments trigger the syllable acquisition process, which is regulated to optimize a criterion function that integrates measures for model pool completeness, orthogonality and stability.

As syllable and word representations are learned incrementally in parallel, Viterbi-decoding is not directly applicable. Therefore, we employ speech unit-spotters to detect already learned syllables/words. Thereby the next more fine-granular representation is employed as background (aka. world-, filler-) model (cf. section 3.6.3). The segmentation of the speech feature stream into syllabic and phonemic background units defines the input to the word layer, which is implemented as a discrete model with syllable symbol sequences modeling the different word units.

For each level of speech granularity a statistical n-gram model is used to complement the acoustic modeling with transitional constraints. This is motivated by findings of ASR-research that statistical language models increase the performance of ASR systems by an order of magnitude. Beside such computational benefits, the main motivation was to integrate mechanisms of statistical learning as observed in infants into our model [Gam05], [Baz01].

Clearly the n-gram models used within our architecture differ from what is used in common approaches to speech recognition. The main difference is the manner in which these models are bootstrapped incrementally in interaction within our architecture. Initial n-grams are chosen to be flat distributions, which are incrementally updated based on the results of the respective unit detection module. This contrasts to statistical languages models as described in section 3.5 which are trained offline using ground truth data comprising up to billions of example-utterances.

6.3 Phones

A phone representation is crucial to make our proposed speech acquisition architecture operative: It allows to convert speech input into sequences of phone symbols. First, this is mandatory to build a phonotactic model of the tutoring language. Second, syllable models can be initialized by concatenating phones HMMs. Next, as shown in fig. 6.7, the syllable spotter requires a phone representation as background model. Finally, a phone representation allows to normalize acoustic scores while recognizing syllables as described in [Kam00].

The phone layer involves three major submodules: First, a solely data-driven clustering module to estimate the phone representation. Second, acquired phones HMMs are employed to convert speech into phone symbol sequences along with segmentation information. This is achieved by means of a Viterbi-Decoder (cf. sec . 3.5). Third, a phonotactic model is estimated from recognized phone sequences. It is implemented by means of two n-gram models which encode phone structure of syllable initial and the final parts respectively.

6.3.1 Unsupervised phone cluster learning

We adopted the method of [Iwa06] (cf. section 5.1) to bootstrap a phone representation. Phones are learned by accumulating a certain amount of speech features. Then, single state HMMs are created using mixtures of Gaussians including 8 component densities as output probability distribution functions. k-Means as described in section 3.1 becomes applied to estimate a probabilistic model of the phone feature space without taking phone transitions into account. Subsequently, frame level counts (cf. [Sha07]) are employed to estimate transition probabilities between these single state HMMs.

To take the time dimension of phones into account a Monte-Carlo-sampling governed by frame level transition statistics is used to determine the most frequent state-sequences. The N most frequent state sequences are concatenated to give phone-models comprising M states linked with Bakis-topology. These initial phone models become further refined using Baum-Welch-training as described in figure 3.6 to give the final phone model pool $\mathcal{M}_P$.

For $M = 1$ phones do not encode time. For $M > 1$, phone models have a temporal dimension of around $10 \cdot M$ milliseconds. This corresponds to a search space with a very high branching factor. To ease decoding, phone model transitions are further constrained with the above mentioned frame level transition probabilities.

Because the number of phone-models is not known a-priori we use the Akaike information criterion (AIC) as described in [Aka74] to optimize the number of phone models:

$$\text{AIC} = k - \ln(L(X|\mathcal{M}_P)) \tag{6.1}$$

AIC maximizes the likelihood L computed on a test-set while penalizing the model complexity which is expressed in terms of the number k of model parameters. Although it would have been possible to further adapt the phone model online, it is kept fixed after training to allow neglecting stability reasons as discussed in section 6.1.5.

Such a phone representation differs from the phoneme-models used in most automatic speech recognition systems: Whereas phonemes are a linguistic concept and refer to minimal meaningful sounds, we think phones to be basic speech sounds without any relation to meaning (cf. sec. 6.1.2). But even if the concepts of phonemes and phones differ, the computational techniques to handle them are almost identical. Thus, our model can profit from all methods described in section 3.6.

6.3.2 Phonotactic Learning

As phonotactics refer to the rules that govern the structure of syllables in a particular language, such rules may be captured in a probabilistic sense. But bootstrapping of phonotactic constraints faces problem that the latter can only be learned from an utterance if the syllable boundaries are known [Chr98]. At first glance, phone sequence as obtained from the phone recognizer do not convey any syllable segmentation markers. However, as discussed in section 6.1.4 each utterance has to be bound by an initial and final syllable. To formalize this idea we complement each phone utterance with *utterance boundary markers.* Given a sequence of phone symbols $[\lambda_P]_1^N(X) = \lambda_P^1, \ldots, \lambda_P^N$ that has been recognized given an utterance X we create an extended sequence by

$$[\lambda_P]_1^N(X) \rightarrow \oplus[\lambda_P]_1^N(X)\oplus \tag{6.2}$$

Hereby, $\oplus$ denotes a boundary symbol which results from the voice activity boundaries. By adding this marker, the boundary becomes more explicit on a symbolic level which enables to capture the boundary constraints probabilistically. For each utterance this approach gives two training segments to capture the phonotactics of the tutoring language: the initial phone-symbols of the syllable at the utterance start and the coda phone-symbols of the syllable at the utterance end. Without an explicit syllable model it is not possible to induce further phonotactically meaningful training segments.

As statistical language models allow to calculate co-occurrence probabilities for any sequence of speech units, they can also provide cues for segmentation [Elm90] [Dav01] [Gam05]. At first glance a natural choice for a phonotactic model would be the phone language model L_P itself. However, as this model encodes the complete transitional phone structure of the tutoring language, we consider a more specific model to be better suited. Furthermore, as L_P encodes the phone sequence in a time directed manner, it is not applicable to utterance initials as for these the n-gram context appears after the boundary symbol.

Here we model the probability for a syllable change by combining two n-gram models P_{SI} and P_{SF} for the syllable initial and final part respectively. Both are estimated from the initial and final parts of the $[\lambda_P]_1^N(X)$ whereby the number of symbols to be taken into account at each boundary is limited by the context size of n. According to the definition of the n-gram model in equation 3.34, P_{SI} needs to be trained by reversing the order of phone symbols in the initial subsequence..

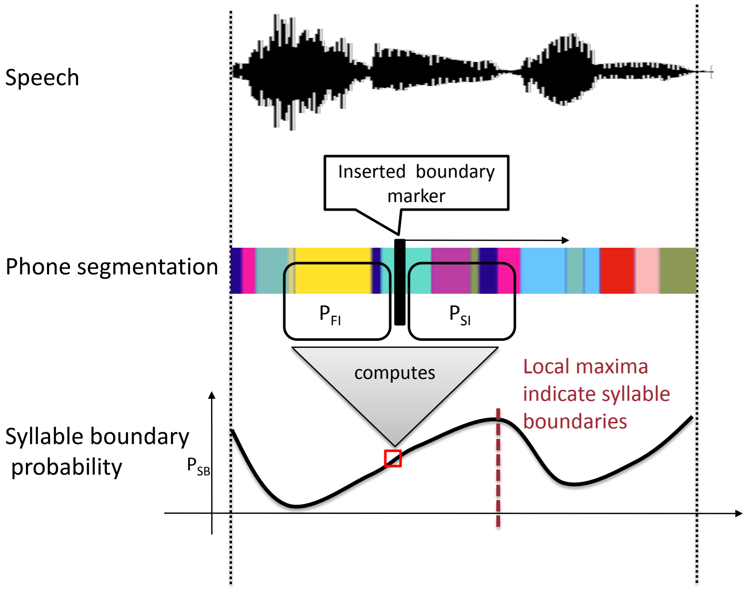


Figure 6.6: Phonotactic parsing applied to a to a detected phone sequence. The boundary marker $\oplus$ is shifted through the phone sequence, which allows to calculate the syllabic boundary probability $P_{SB}(k)$ for each position k. The resulting syllable boundary probability function P_{SB} can be used to decompose a given speech segment into syllabic sub-segments, which is a prerequisite for developmentally plausible syllable learning.

The context size of the two n-gram models needs to be an appropriate trade-off between discriminative power, trainability and applicability given the task to learn syllabic structure constraints. Whereas a uni-gram model would only take the first and the last phone symbol of adjacent syllables into account, too large context-size might exceed the actual syllable length. Therefore a bi- or tri-gram model seems to be an appropriate trade-off between discriminative power, trainability and applicability. To cope with unobserved phone-sequences, Katz-smoothing was integrated (cf. [Kat87] and section 3.5).

Phonotactic parsing

As discussed in section 6.1.4 the purpose of the phonotactic model is to implement a parsing mechanism to determine the number of syllables contained in a speech segment X. The first step is to compute a syllable boundary probability function $P_{SB}(k|X)$ based on the sequence of recognized phones $[\lambda_P]_1^N$ corresponding to X. The probability for a syllable change after the phone symbol k is computed as the product of P_{SI} and final P_{SF} by splitting the argument phone sequence after phone k and extending both sub-sequences with the boundary marker $\oplus$ accordingly:

$$P_{SB}(k|[\lambda_P]_1^N) = P_{SF}([\lambda_P]_1^k\oplus) \cdot P_{SI}(\oplus[\lambda_P]_{k+1}^N) \tag{6.3}$$

Thus, the complete function can be obtained by shifting a boundary marker through the phone sequence and to evaluate P_{SB} for each current boundary marker insertion point. Ideally, the resulting function peaks at the syllable boundaries. As noise is likely to obfuscate boundary peaks, a subsequent low-pass filtering is used to smooth the signal. The complete process is visualized in figure 6.3.2.

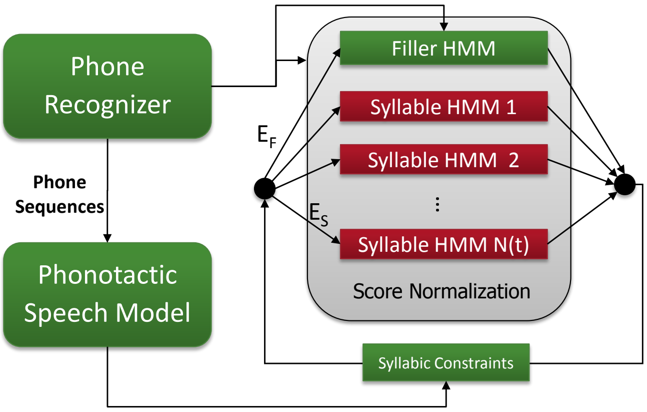


Figure 6.7: The syllable spotter implementation. The phone model pool $\mathcal{M}_P$ is employed as filler model and to normalize acoustic scores. Learned phonotactics further constrain the Viterbi decoding. Segments that contain instances of not yet acquired syllables will be matched by the generic phone model. E_F and E_S denote filler insertion and syllable insertion penalty respectively. Beside language model probabilities, both are commonly employed to further constrain model transitions in speech unit spotting systems (cf. section 3.6.3)

6.4 Syllables

There are two main conceptual approaches to represent syllables in a computational model. First, sequences of phone symbols or probability distribution over such sequences could be employed as syllable models [Mar95] [Roy00] [Bal03]. Second, syllables could be modeled as distinguished entities [Mur04]. We favor the latter approach, mainly because it is less prone to co-articulatory effects as discussed in section 6.1.5. More precisely, we model syllables as continuous-density HMMs with a fixed number of states, Bakis topology and diagonal covariances matrices in the component densities.

6.4.1 Syllable spotting

Initially the syllable representation does not contain any models. Incoming speech is analyzed solely by the phone-recognizer and the voice activity detector as depicted in figure 6.5. After phones have been learned, the syllable spotting process is started, with the phone models $\mathcal{M}_P$ embedded as background-model. Newly acquired syllables are glued dynamically into the search space as highlighted in figure 6.7. As usual for speech unit spotting systems, the entry into the background model is penalized. Both, phone and syllable transitions are constrained by the syllable transition model L_S . In contrast to other speech processing frameworks like HTK [You06] or ESMERALDA [Fin03], we decode only speech activity segments as provided by the voice activity tracker, and thus do not require an acoustic silence model.

The syllable spotter converts each utterance into a non-overlapping sequence of phone and/or syllable segments. These utterance hypotheses trigger a variety of sub-processes as (partially) depicted in figure 6.5:

- Subtraction based training segment generation as described in section 6.4.2
- Incremental learning of the syllable language model as discussed in section 6.4.5
- Word learning and spotting as outlined below in section 6.5

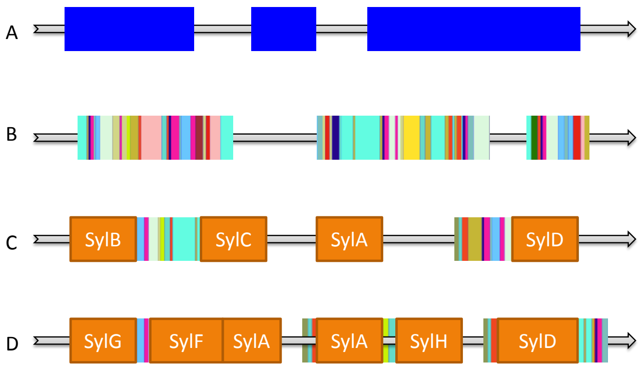


Figure 6.8: Decoding results of the syllable spotter at different time instances. Initially (A) only voice activity segments indicated by dark blue are detected. Learned phones start the actual syllable spotting process (B) whereby each phone has been assigned a different color to ease visual inspection of the segmentation results. Even if no syllables are present initially, the background phone model allows a complete segmentation of the input signal. Learned syllables (C) (here arbitrarily named) give a segmentation of VA-framed utterances into parts already represented by the syllable model and not yet syllabificable phone-sequences. After the syllable representation has converged (D), the complete speech input is segmented into syllable segments. But as speech unit spotting techniques are not yet mature enough to result into completely matching segmentation, boundary phone artifacts are unavoidable when evaluating the model on actual speech data.

- Different regulatory processes that modulate the syllable bootstrapping process as delineated in section 6.4.4.

Some of these processes rely on confidence scores for each partial segment-hypothesis. Thus, partial path likelihoods are normalized to give acoustic confidence values as described in section 3.6.4. More precisely, this is done according to equation 3.36 by normalizing partial path likelihoods by partial segment scores as obtained from the phone recognizer model for the respective segments.

Figure 6.8 depicts the different processing stages. After the initial phase A, where only the voice activity detection takes place because syllable and phone representations are still empty, phones provide a segmentation into phone-segments in phase B. Emerging syllable models will increasingly contribute to the segmentation of the speech signal as shown in sub-figure C. Finally, after the syllable representation of the input language has been converged, syllable segments almost completely cover the utterance segments as depicted in sub-figure D.

6.4.2 Training segment generation

The syllable learning in our framework relies on training segments that are clustered to reveal the syllabic structure of the tutoring language. Thus, a central mechanism of our architecture is the extraction of training segments from the speech signal. To make such a clustering process functional, each segment must be ensured to contain one and only one syllable segment. According to our above discussions, length is not a reliable cue to decide what a syllable is [Bra08]. Hence, we rely on the phonotactic parser as detailed in section 6.3.2 to extract reasonable training segments. By doing so we couple the structural constraints captured by the phonotactic model with the syllable learning process.

A first computational strategy to generate training segments, that follows the way infants seem to learn the syllable structure of their parental language is to use only segments that contain with high (phonotactic) confidence one single syllable. As discussed in chapter 2, this applies to some extent to infant directed speech, so this assumption seems valid when developing a system for speech structure acquisition. Furthermore, isolated word-utterances can be considered to significantly increase the robustness of the clustering process in the initial bootstrapping phase, as with shorter utterances instable syllable models are less likely to propagate segmentation errors into the training segments.

However, as discussed in chapter 2, not all cultures match the tutoring speech to their infant's perceptual abilities. This requires infants to extract syllabic training segments mainly from continuous speech. Hence, we propose to use a more elaborate scheme that aims to cope also with such situations. It is motivated by our findings in section 6.1.4, that training segments can be extracted from continuous speech using the principle of subtraction. We consider three different mechanisms to obtain syllable training segments from the speech signal:

1. Segments as indicated by the phonotactic parser. These are pruned (too short ones like erroneous inserted phone artifacts) or optionally split (multi-syllable) according to the syllable boundary function P_{SB}. The resulting speech snippets can than be assumed which high confidence to comprise just a single syllable.
2. Segments that are obtained by applying the principle of subtraction to syllable spotting results followed by a phonotactic test about mono-syllabicity.
3. Syllable segments as detected by the syllable spotter followed by a phonotactic test about mono-syllabicity.

The resulting segments are used to trigger the clustering process. Even if all three cases are closely related, they differ with respect to robustness: Generic syllable models that emerge in the initial bootstrapping stage are unlikely to give reliable syllable segments. However, as it is not clear to which amount these mechanisms could complement each other, we consider all for evaluation. By implementing these mechanisms into a common framework we hope to reveal which mechanisms are mandatory for early infant speech structure acquisition.

6.4.3 Incremental clustering of syllable segments

The syllable learning approach proposed in this section embeds seamlessly into the general notion of clustering schemes as discussed in section 3.1. It implements an incremental divisive clustering scheme that models syllables as they appear in time. Technically it is related to some extent similar to Leader-Follower clustering described in section 3.2. Figure 6.9 summarizes its computational realization. Like Leader-Follower clustering, the approach implements two important steps. First, the syllable model that matches best to the training segment X is detected. Second, either a new syllable cluster becomes created or the best matching model becomes adapted in direction to X. The former applies if the novelty of the training segment exceeds the novelty threshold θ.

Newly created syllable models become dynamically integrated into the syllable spotting process as depicted in figure 6.7. By doing so, the system is equipped with the ability to extract training segments for increasingly complex speech input. Adaption of existing syllable models makes those to drift in the syllable model space in direction of the training samples. This we assume to give an asymptotically stable and discriminative syllable representation.

Let X a new training segment, $\theta \in \mathbb{R}_+$ a novelty threshold, and $\mathcal{M}_S$ the set of already acquired syllable models. X becomes processed as follows:

1. Determine the model λ_S^* which is most likely to explain the given segment X in terms of maximum likelihood

$$\lambda_S^* = \arg \max_{\lambda:\mathcal{M}_S} P(\lambda|X) \tag{6.4}$$

2. Consider two cases depending on the segment novelty $\nu(\lambda_S^*, X)$
   - (a) $\nu(\lambda^*, X) < \theta$: X is assumed to contain a not yet represented syllable. Create a new model λ_S^{NEW} for X and add it to $\mathcal{M}_S$.
   - (b) $\nu(\lambda^*, X) \geq \theta$: The model λ^* seems to be appropriate to model the current segment $X \rightarrow$ Adapt λ_S^* with X.
3. In case of (a) initialize $F(\lambda_S^{\mathrm{NEW}})$ with the N best matching training samples of λ_S^*. Adapt $F(\lambda|X)$ with $P(\lambda_S^*|X)$ in case of (b).

Figure 6.9: The syllable clustering algorithm. Triggered by syllable training segments, the approach can be characterized as a Leader-Follower clustering scheme with a speech confidence based history model as novelty function.

Novelty detection

The first step to improve a syllable representation $\mathcal{M}_S$ given a training segment X is to match X against $\mathcal{M}_S$ as this allows to determine the novelty of X with respect to the existing representation. Comparable to mixture density evaluation, this requires to compute the posterior probability $P(\lambda_S|X)$ for each $\lambda_S \in \mathcal{M}_S$. Computationally these are the confidence values calculated according to equation 3.36. However, as these values are not bound to $[0, 1]$ and depend on the acoustic structure of the current syllable, it is necessary to map them to a domain in which the novelty threshold θ can be applied [1].

We define a novelty function $\nu(X, \lambda_S)$ as follows. For each syllable model λ_S we accumulate the posterior probabilities of former training segments in an histogram $H(\lambda_S)$. This is approximated by a probability distribution with the density $f_{\lambda_S}(p)$ using a Parzen window model (cf. section 3.2.3). To calculate $\nu(X, \lambda_S)$ in algorithm 6.9 the corresponding cumulative distribution function $F_{\lambda_S^*}$ is employed to map $P(\lambda_S^*|X)$ onto the respective quantile value:

$$\nu(\lambda_S, X) = F_{\lambda_S}(P(\lambda_S|X)) = \int_{-\infty}^{P(\lambda_S|X)} f_{\lambda_S}(p)dp \tag{6.5}$$

The decision threshold θ needs to be chosen heuristically as common for all unsupervised incremental clustering techniques. This is considerably simplified as the novelty values are distributed in $[0, 1]$. Intuitively, the value of θ thereby modulates the granularity of the emerging syllable representation.

[1] Conversely formulated, θ needs to be adapted to the current model. But this would give a double-adaption of θ in combination with the regulatory adaption of θ described in section 6.4.4. Thus, we keep the notion of novelty scale adaption for sake of readability.

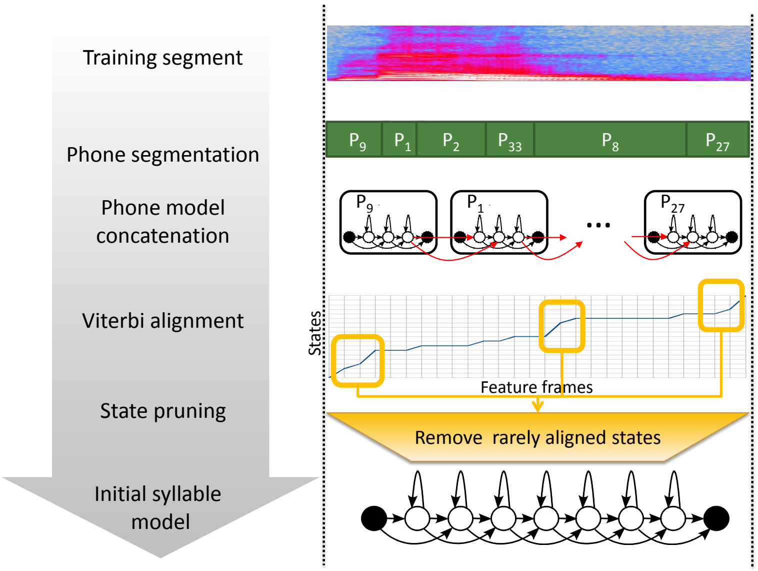


Figure 6.10: The initialization process of syllable models. The phone segmentation gives a sequence of phone models to be concatenated. To avoid artifact phones to be included into the initialization model, a topology optimization step is introduced. It attempts to prune states of the merged HMM that contribute little to the Viterbi decoding of X.

Syllable model estimation

Algorithm 6.9 leaves two implementational details open. These are the model initialization and adaption. Even if this is beneficial from a design point of view, specific solutions have to be chosen to evaluate the model.

Model initialization

As discussed in section 3.6, HMM parameter estimation crucially depends on a proper model initialization. When the segment novelty decision indicates the creation a new model, a straight forward approach would be to clone the best matching λ_S^* to create an initialization model. But this encompasses two problems. First, it leaves open how to create the first syllable model. Second, syllable models derived during the initial bootstrapping phase would rather represent generic than actual syllables. Thus, we favor a different approach which is summarized in figure 6.10.

Syllable models are initialized by concatenating the best matching phone-sequence $\lambda_P^1, \ldots \lambda_P^N$ for a training segment X.

$$\lambda_S^{\text{NEW}} = \lambda_P^1 \circ_B \lambda_P^2 \circ_B \ldots \circ_B \lambda_P^N \tag{6.6}$$

Here $\circ_B$ denotes a concatenation operator that merges two (phone) HMM models into a single HMM with Bakis-topology.

To keep the syllable representation consistent with respect to model complexity, we further post-process the initialization model λ_S^{NEW} to limit its number of states. Prior experiments showed that even with high penalties for phone-to-phone transitions the number of phones per syllable segment differs considerable between syllable models without any clear phonological reason. Therefore we calculate a Viterbi-alignment between the features X to the states in λ_S^{NEW} and prune the states which contribute less than a certain ratio υ to the alignment. By applying a state-wise pruning, different numbers of states per syllable model are still permissible depending on the acoustic complexity of the syllable.

Model adaption

The adaption case (2b) in algorithm 6.9 is realized by a single MAP-training iteration. As incremental MAP-adaption rapidly leads to model degradation as shown in [Bra05], we propose to use a windowed buffering scheme for MAP-adaption: for each syllable model λ_S the system keeps a history $\mathbf{X} = X_1, X_2, \ldots, X_{N-1}$ of past training segments. Given a new update segment X_N we update the respective model using all statistics gained from the segments in $\mathbf{X} \bigcap X_N$ (cf. section 3.4.2).

As the number of segments contained in $\mathbf{X}$ increases, the inherent statistics about the parameterization of the respective syllable model become more reliable. Thus, to further increase the model quality, syllable models are re-estimated using iterative BW-training with $\mathbf{X}$ as training sample. More precisely we use a simplified variant, where the training samples are split into equally sized portions assigned to each state (cf. [Fin03, sec. 5.7.2] for a related approach). Thus, BW reduces to an iterative EM-refinement of each state distribution. This simplified approach was chosen because of much lower computational demands. Specifically this was necessary to avoid delay times when interacting with the system. Furthermore, preceding experiments have shown no significant performance drawback of this approach compared to full BW-training.

The EM-retraining allows to replace the adapted model with its EM-reestimate. Subsequently, all elements in $\mathbf{X}$ are discarded, as this is necessary to ensure the scalability of the system. Otherwise, the memory demands of the system would linearly grow with the number of processed training segments. Models that have been re-estimated at least once using ML-training are referred to as *stable*. Because of the dominant effect of the state data likelihoods, transition probabilities were chosen to be fixed during training.

6.4.4 Regulation

As discussed above in section 6.1.4 plasticity and stability have to be modulated to make an unsupervised learning approach to reveal a convergent and robust description of the data in terms of clusters. Clearly, such modulations are necessary especially for incremental learning methods like the proposed syllable clustering scheme. Hence, we embed different modulatory loops into our architecture.

Global Control

Global control links the system behavior in a top-down manner to control parameters that modulate the clustering process. This requires to calculate a set of regularization terms commonly used for unsupervised learning tasks: completeness Γ, orthogonality η and stability ψ. We realized these terms as follows:

Model spotting coverage $\Gamma(t)$ measures the completeness of the representation at a given time. It is defined as the ratio of accumulated syllable-segment lengths to the overall amount of speech.

Model co-activity measures the mutual dependence between all syllable models at a specific time. Optimally syllable models are orthogonal with respect to their discriminative power, i.e. only one model is active at a time. It is measured pairwise by comparing the relative confidence rankings of models on a set of benchmark segments (e.g. a sliding window over the training segments).

Intuitively, models that match to the same segments can be considered to be correlated. For two models i and j the model co-activity is denoted with $\eta(\lambda_i, \lambda_j, t)$.

Pool stability $\psi(t)$ is defined as the ratio of stable models to the non-stable models (cf. sec. 6.4.3).

To compute Γ and η a history interval needs to be defined. All three measures are distributed in $[0, 1]$ by definition. This allows to rephrase the incremental syllable clustering problem as an optimization problem.

$$\Gamma + \psi - |\eta| \rightarrow \text{max!} \tag{6.7}$$

Thereby, $|\bullet|$ denotes a common matrix norm. Intuitively, this regularization model drives the system to establish homeostasis as soon as the syllable representation allows to completely model the input speech. Because it is not possible to find an analytic solution for this problem, we propose two heuristics which attempt to maximize this criterion function.

(I) First, the creation of new models is modulated by the pool stability. New models are created only if

$$\psi(t) > \Gamma(t) \tag{6.8}$$

Otherwise the best pool model becomes updated. According to this heuristic, the creation of new models is eased if speech coverage is low. Vice versa it prevents the creation of new models if $\mathcal{M}_S$ is already suitable to model the speech input.

(II) Whereas the default acquisition loop assumes $\nu(\lambda^*, X)$ to be greater than a fixed threshold it might be more appropriate to use an adaptive threshold. Such a threshold can be chosen by:

$$\theta = \theta_0 \cdot (1 + \beta \cdot \psi) \tag{6.9}$$

This heuristic is inspired by the idea to ease the creation of new models if the stability of $\mathcal{M}_S$ is high. Vice versa, low stability prevents the creation of new models. Thereby β defines a weighting factor.

Local control

Learning via repetition as often observed in parent-infant interaction is hard to realize by using the clustering method 6.9. This is because recently created syllable models rely to large amounts on the inherited phone model parameterization and only partially on the syllable training statistics. To overcome this problem we introduce a local stability criterion into the clustering process. The novelty threshold θ is further adapted depending on the best matching model λ^*_S.

$$\theta \rightarrow \theta(\lambda^*_S) = \theta_0 \cdot (1 - \exp(|H(\lambda^*_S)|) \tag{6.10}$$

$|H(\lambda^*_S)|$ denotes the number of segments in the training history of λ^*_S (cf. section 6.4.3). This scheme complements the global learning control mechanisms depicted above. It attempts to modulate the confidence of novelty estimates depending on the amount of information encoded in the training history. It ensures that only syllable models that have been estimated on a sufficiently large set of training sample can cause the creation of new syllable clusters. Thus, it can be considered to temporarily reduce local plasticity in favor of stability.

6.4.5 Syllable transition modeling

As for phones, syllable spotting results are used to incrementally estimate a bi-gram transition model L_S. To increase the robustness of the learning process, only utterances which are segmented almost completely into syllables without dominating phone sub-segments, are considered as training samples. This is motivated by the idea, that utterances which can not be represented in terms of syllables are rather bad examples to learn statistical constraints about syllabic transition-constraints. Thus, the estimation of L_S is delayed automatically as long as the syllable representation has not been converged.

6.5 Words

In section 6.1.5 we motivated why words can be modeled as syllables sequences. More precisely, each word W is modeled as a sequence of syllable symbols:

$$\lambda_W = \lambda_S^1, \lambda_S^2, \ldots, \lambda_S^N, \qquad \lambda_S^i \in \mathcal{M}_S \tag{6.11}$$

This contrasts the way syllables are modeled, as words are not concatenated syllable HMMs but syllable symbol tuples. This has severe consequences for the implementation of the word layer depicted in figure 6.5. Most notable, the feature space of the word layer differs from the acoustic speech feature space as present for phone and syllable layer. The word layer employs the recognition results of the syllable layer as input signal. Hence, word models need to be modeled as discrete HMMs with the number of states being equal to the number of syllable symbols. But as neither HMM decoding nor search space compilation depend on the type of used OPDFs, the general setup of the word layer is similar to what has been proposed for syllables. Whereas the phone-representation is used as a generic background model for the syllable layer, the word layer employs the set of already acquired syllables (or more precisely the respective syllable symbols) to model not yet represented words.

As already highlighted in figure 6.8, the syllable spotter is unlikely to give a perfect segmentation. Artifact phones will be occasionally inserted. Not yet represented syllables will be matched by the background phone model. For sake of simplicity we restrict the word layer input to syllable spotting result sequences that do not contain any significant portion of background phones. This is implemented by pruning artifact phones using a simple length threshold, prior to entering the word layer. Second, utterances that are not syllabified completely are discarded for lexical learning. Whereas the latter pruning rule is not strictly necessary for our system to function, we consider it to improve the stability of the lexical model being learned: As long as the syllable representation is not sufficiently comprehensive to cope with an utterance completely, it seems is unreasonable to acquire new words from the resulting mix of syllable segments and not yet-syllabified phone subsequences.

The lexical acquisition mechanism which allows to bootstrap a lexicon λ_W is depicted in figure 6.11. Starting with an empty lexicon we combine algebraic learning (cf. [Gam05]) with co-occurrence based word acquisition (cf. [Asl98]).

The algorithm involves three steps. The first one is inspired by the finding that CDS comprises a large ratio of mono-syllabic words utterances. Thus, (partial) utterances that contain only one

Let $S = \lambda_S^1 \lambda_S^2 \dots \lambda_S^N$ a sequence of syllables, and $\mathcal{M}_W$ the set of already acquired word models.

1. **if(**$S \in L$**)** return, because words are modeled as discrete symbol sequences, and a syllable sequence which is already in $\mathcal{M}_W$ does not need to be re-added.
2. **if(**$|S| == 1$**)** add S to $\mathcal{M}_W$ because every isolated syllable is a word.
3. **else if(**$\mathrm{L}_S(S) > \Theta$**)** add S to $\mathcal{M}_W$ because the syllables in S co-occur with such a high probability that it is reasonable to assume S to be a word.
4. **else** Match $\mathcal{M}_W$ against S using $\mathcal{M}_S$ as background model:

   $$S \to \hat{S} = [\lambda_1, \dots \lambda_N], \lambda_i \in \mathcal{M}_W \cap \mathcal{M}_S \qquad (6.12)$$

   Extract partial utterances by splitting $\hat{S}$ on each word symbol. Apply algorithm 6.11 recursively to each resulting syllable sequence $S'_1, \dots, S'_K$ which is composed by elements of $\mathcal{M}_S$.

Figure 6.11: The lexical learning algorithm.

single syllable are added as words unconditionally if they are not already contained in the lexicon $\mathcal{M}_W$. Second, we aim to exploit co-occurrence patterns between adjacent syllables to detect new words: Polysyllabic words are acquired as indicated by the co-occurrence probability of their syllable constituents. The co-occurrence threshold Θ modulates the sensitivity of the algorithm. Finally, we apply the principle of subtraction to decompose the sequence S into existing word phrases and residual background-model sequences. Technically this is implemented by applying the word spotter locally to S. The resulting segmentation of S reveals a non-empty set of residual segments. Those are the syllabic background segments in the segmentation result. This residual approach allows to reveal words that are not tangible given S itself. Residual segments are not treated as new words models automatically, but become recursively processed by the algorithm.

To make algorithm 6.11 to reveal poly-syllabic words, a careful choice of Θ is crucial. A fixed value is not applicable here, as n-gram probabilities always depend on the size of the n-gram vocabulary. Thus, we propose to adapt Θ dynamically depending on the number of already acquired syllable models:

$$\Theta = \frac{\Theta_0}{|\mathcal{M}_S|^{\min\{n_{\mathrm{L}_S}, |S|\}}} \cdot d(t) \qquad (6.13)$$

The basic idea of this formula is to normalize a heuristically chosen base threshold Θ_0 with the number of possible argument sequences. Basically this depends on the context-size n_{L_S} of L_S. But as the length $|S|$ of S might be shorter than the context, the minimum of the actual length and the context size of L_S is used to determine the normalization term. The last term $d(t) = \max(1, d_0/(t+1))$ denotes a delay function with the number of processed utterances t as argument. Initially when L_S has not yet converged, it increases Θ to diminish the risk to acquire incorrect poly-syllabic word models. As L_S has converged - controlled by a fixed time-offset d_0 - , the value of $d(t)$ simplifies to 1.

This approach to lexical learning contrasts to the model of Gambell and Yang proposed in [Gam05] (cf. sec. 4.3), as it relies on a self-referential decomposition - implemented as word spotting with the syllables as filler-model - of the input signal to reveal segments for lexical bootstrapping, and not a direct dictionary-lookup. Thus, our proposed model takes not only the local context into account when learning new words, but the complete utterance.

However, it is clear to us that our approach is likely to be less powerful compared to the symbolic lexical learning mechanism of Gambell and Yang. This is because the latter additionally relies on stress information. This is not applicable here because of missing stress cues. Language *dependent* stress detection has been realized by using heuristically designed rule-systems to detect pitch and duration patterns [Sat03]. But as discussed in section 2.2.2, stress is highly language dependent, and its properties are not assumed to be innate. It is therefore not surprising that no language independent signal processing techniques for stress detection have been proposed yet.

6.5.1 Top-down error correction

Even if not intended while designing the system (and actually first considered as an error while debugging it), the proposed architecture for word recognition turned out to implement a local error correction scheme for poly-syllabic words. As syllables define the input for the word-spotter, it is intuitive to suppose syllable recognition errors to propagate into the word recognition results. But as words are the result of HMM-decoding on the syllable input sequence, the cost-minimal path through the search lattice will be (at least in some situations and given an appropriate insertion-penalty configuration) correct an incomplete syllable input into the correct word sequence.

For instance, let the utterance `take this blue cup` be sampled from the artificial language 7.15. An incompletely recognized syllable sequence `take blue cup`, is not decomposable into words as `take` always co-occurs with `this`, and `take this` will have been learned as bi-syllabic word. Thus, the word parser is likely to complement the input to give the correct word sequence `{take this} {blue} {cup}` as this is the cost-minimal solution. From a bio-inspired point of view, such a property is highly desirable, as it shows how higher-level knowledge can provide a modulatory feedback to correct an erroneous input signal.

6.5.2 Basic syntax learning

Similarly to syllables and phones, the transition probabilities of adjacent words are captured by an incrementally trained n-gram model L_W. Whereas statistical learning on phone and syllable level was necessary to bootstrap the next higher-level speech representation, our framework does not employ a similar scheme on word level. The estimation of L_W is rather beneficial when evaluating the proposed lexical learning scheme in section 7.5.1 and aims to further extensions of our model as delineated in section 9.1. Nevertheless, the transitional structure is necessary to disambiguate word decoding under certain conditions.

6.6 Scientific contribution

This chapter proposed a computational model for unsupervised speech structure acquisition. The underlying principles are motivated mainly by findings from developmental psychology discussed in chapter 2. The technical backend of our architecture are speech processing techniques common to ASR systems. More precisely we used the decoder, search space compilers and a feature front-end provided by the Sphinx4 ASR system presented in [Wal04]. Many aspects of our model are inspired by the works as discussed in chapters 5 and 4. Most notably we followed the ideas of [Dav01], [Iwa04] and [Kit03], while designing the model. The main contribution of our model is that it combines the unsupervised incremental learning of phones, syllables and words into a unified architecture. In contrast to previous works, we focused on scalability, acoustic speech as sole input, incremental learning, developmentally plausible processing schemes, and language independence. More precisely, our model contributes to field of computational speech structure acquisition in several ways:

First, our model emphasizes on processing principles (cf. section 6.1.4) that are developmentally plausible. Previous models on acoustic speech acquisition, mostly neglected this premise. Our system design is focused on making as realistic assumptions as possible concerning the processing mechanisms and sources of information. Previous computational models often over-estimated innate abilities of young infants. This especially includes online processing, which was highlighted in [Gam05] to be an often neglected aspect when building models for speech acquisition. For instance, Brent and Cartwright introduced a model for lexical acquisition in [Bre96] that optimizes a word lexicon by maximizing a metric that is computed over a complete test-corpus. However, infants are unlikely to implement such approaches to reveal the structure of speech.

Next, our model is completely unsupervised and does not rely on a arguable innate speech representation. This contrasts to many previous models. For instance, innate phoneme-recognition abilities were assumed by [Roy00], a structurally constrained syntax model was supposed by [Iwa06], or phonotactic knowledge was assumed to be innate in most works synopsized in chapter 4. In contrast we implement a layered, partially hierarchical model comprising coupled speech representations for different levels of speech granularity. Thereby, no innate constraints have been built into the proposed architecture that would limit the application to a particular language, except the idea that words are organized in terms of syllables. In contrast, most models as presented in chapter 4 attempt to reveal structure on a single granularity level.

Third, inspired by the work of Iwahashi (cf. sec. 5.1) we propose a partially new method to bootstrap a phone representation. Furthermore, and not yet proposed in the field of acoustic segmentation, we propose how to learn phonotactics directly from an unconstrained acoustic speech input. Although such a step is mandatory to model human speech acquisition, almost all other previous models on speech acquisition have neglected this aspect. Our model learns structural restrictions on what makes well formed syllables in the tutoring language in a solely data-driven unsupervised manner.

Fourth, our model is online-capable and scalable as it is based on HMMs for speech representation and thus allows to use state-of-the-art decoding techniques. This contrasts to many works outlined in chapter 5, which rely on computational approaches that have not been shown yet to scale up to the complexity of human language. Furthermore, the complete architecture is implemented in a way that allows further integration into an embodied infrastructure (cf. chapter 8).

Next, our model relies on developmentally plausible processing schemes, integrated in a way that has not yet been proposed for acoustic speech structure acquisition. Our model is the first attempt to use self-referential bootstrapping principles like subtraction learning for speech unit extraction. Furthermore, our model presents the first attempt to apply self-referential learning on different layers of speech granularity within a single architecture. By applying the principle of subtraction for syllable and word learning, we take into account what developmental psychologists have revealed from infant development (cf. chapter 2). It implements a phonotactic parsing approach to determine the syllabicity of speech segments. This has been proposed previously for the symbolic domain only. However, as symbolic speech processing disregards many difficulties that have to be addressed for acoustic speech, we consider our work to be a substantial contribution to the field of unsupervised phonotactic learning.

Finally, we propose a set of local and global regulatory control schemes to modulate the unsupervised syllable clustering process. In contrast, most existing clustering schemes lack of such regulative means and focus on heuristically determined system parameters. However, given the complexity of speech, we consider the former to be mandatory for data-driven incremental speech structure acquisition.

6.6.1 What this model is not

The proposed model focuses on high level perceptual functions and not on neural mechanisms underlying speech perception. Even if both are likely to be depend on each other, we consider a purely functional view nevertheless enlightening with respect to computational mechanisms and principles underlying the processes of speech acquisition. It is out of scope of this work to organize speech acquisition and perception in terms of brain-like structures.

In this chapter we focused solely on perceptually driven speech acquisition. As argued above this will give at best a set of perceptual units similar to what humans perceive as words. However, in a strict sense the acquired word models are not symbols because they do not refer to something. Meaning has to be derived from embodied interaction, which requires a model for speech acquisition to be embodied. As mentioned in chapter 2 such an embodiment may be even crucial to acquire perceptual speech units. Therefore, we will describe first attempts toward grounded speech acquisition in chapter 8. There we present how to embed the proposed system into an embodied agent.

CHAPTER 7

Evaluation

As estimated by [Har95], approximately 10 million words per year are addressed to young infants until the age of five. However, even if there are vast collections of audio files available, only few corpora are suitable to address the problem of speech structure acquisition. The huge corpora used to train ASR system are partially well annotated, but lack of the properties of child-directed speech. Corpora that address infant-parent interaction like MOTIONESE [Roh04] mostly contain only small samples per speaker or are often only automatically annotated.

For our evaluation we focus on the aspects summarized in section 6.6. The basic layout of the last chapter was kept for evaluation: we evaluate the three layers per se under various conditions and parameterizations. This enables to validate the function and performance of the different sub-processes. Furthermore, we evaluate how the complete system performs under realistic conditions using a large CDS-similar read speech corpus as input.

7.1 Performance Metrics

Speech recognition performance is commonly assessed by measuring three types of errors [Hua01, sec. 9.2].

- Substitution: a correct word is being substituted by an incorrect word
- Deletion: a correct word is missing in the recognition hypothesis
- Insertion: a supernumerary word is contained in the recognized word sentence

These errors are commonly expressed in terms of *word error rate* (WER), which is computed by

$$\mathrm{WER} = \frac{\textit{insertions} + \textit{deletions} + \textit{substitutions}}{\#\ \textit{words in test utterance}} \tag{7.1}$$

To compute the WER for a recognition hypothesis, it needs to be matched against the correct utterance. This problem is referred to as *maximum substring matching.* It can be solved using dynamic programming techniques [Hua01] [The03]. To apply maximum substring matching, it is necessary that detected models are labeled with the same naming scheme as used in the corpus annotation.

As our experiments will rely on a limited set of test samples only, it is necessary to assess the statistic significance of the results. Therefore we adopt the *BootLog* approach proposed in [Bis04] to compute confidence intervals for our results where possible. Thereby, random subsets of the test-results are sampled with repetition to obtain a frequency distribution of the statistic in question.

7.1.1 Model Labeling

Clearly, a WER cannot be calculated if an unsupervised learning approach has revealed the speech representation. This is because models will not follow any naming scheme and will be rather indexed arbitrarily (mostly in order of creation). One approach to overcome this problem, is to label all models according to the mode of their discriminative function [Mur04].

To ensure the training of meaningful syllable models it is necessary to assess the system behavior when assigning training segments to emerging models in alg. 6.9. This is only possible by using additional supervised information about these training segments. Given such an annotation of the speech signal, a *training confusion matrix* T_{conf} can be obtained by applying the following optimization procedure

$$\begin{aligned} T(i,j) &= \#\text{ segments with label } l_i \text{ which} \\ &\quad\ \text{were used to train model } \lambda_j \\ \phi_{\max} &= \arg\max_{\phi\in\Phi} \sum \operatorname{tr}(\phi(T)) \\ T_{\text{conf}} &= \phi_{\max}(T) \end{aligned} \tag{7.2}$$

This schemes implements an implicit model labeling that reorders models to maximize the trace of the matrix. Thereby, Φ denotes the set of all possible column permutations.

As opposed to supervised machine learning tasks, the number of models M does not necessarily equal the number of classes C. Thus, to make the optimization scheme functional, T needs to be extended with dummy columns if there are less models than annotation labels.

To ease comparative evaluations a scalar statistic can be derived from T_{conf} by calculating the *training confusion ratio* t_{conf}:

$$t_{\text{conf}} = \frac{\operatorname{tr}(T_{\text{conf}})}{\sum\limits_{i,j} T_{\text{conf}}} \tag{7.3}$$

The ratio denotes the average training confusion. It tends to be 1 if training segments with the same label are assigned to only one model for training.

By applying the method 7.2 to the keyword spotting results, a similar statistic can be obtained, which we refer to as *detection confusion matrix* D_{conf} and *detection confusion ratio* d_{conf} respectively. To calculate these statistics, a sliding time window needs to be defined, which functions as accumulator of the system properties.

7.1.2 Segmentation quality

The most basic metric to assess segmentation quality is to calculate the ratio of the number of detected boundaries against the number of ground truth boundary markers. Let S_d and S_e the detected and the expected number of segmentation markers. By computing $S_d - S_e$ a simple metric can be derived that indicates over and under-segmentation but lacks of any kind of normalization [Ave01]. A derived measure that applies a basic normalization is $100 \cdot S_d/(S_e - 1)$ [Pet96].

What is an optimal segmentation of an acoustic signal? A generic framework for evaluation has been proposed by [Qia08], which combines different criteria for optimality. The number of segments gives a first hint about segmentation performance. However, as this does not take the position of segmentation markers into account, this metric is not rich enough to assess how a segmentation algorithm performs on a data set. A method is just required to find $S_d = S_e$ markers to reach the global performance optimum. Therefore, performance of segmentation methods is often assessed indirectly, for instance by measuring WER of a recognizer that setups on an obtained segmentation. However, most reliable performance measures are obtained by comparing segmentation results against a gold standard segmentation like a manually segmented speech corpus.

Symbolic speech segmentation performance can be evaluated using methods from information retrieval. To account for over- and under-segmentation *precision* $\mathcal{P}$ and *recall* $\mathcal{R}$ (also referred to as *prestational index* in [Ave01], or *sensitivity*) are calculated respectively by

$$\mathcal{P} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{7.4}$$

$$\mathcal{R} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{7.5}$$

A *false-positive* result correlates to a situation when the system has detected a non-existing segment, whereas a *false-negative* indicates that it has failed to detect a segment. Whereas it is easy to optimize one of both measures, good segmentation quality requires both to become maximized. Therefore, precision and recall are often condensed into a single metric called *F-measure.* It is computed as the weighted geometric mean using a weighting parameter α, that modulates the relative importance of $\mathcal{P}$ and $\mathcal{R}$:

$$F_\alpha = \frac{(1+\alpha)\mathcal{P}\mathcal{R}}{\alpha\mathcal{P} + \mathcal{R}} \tag{7.6}$$

Commonly α is chosen to be 1. For instance, given the utterance `littleyellowduck` a segmentation into `little yellowduck` yields a word precision of 1/2 (`little` out of `little` and `yellowduck`), and a recall of 1/3 (`little` out of `little`, `yellow` and `duck`), which combines to $F_1 = 0.4$ (cf. [Gam05]). Occasionally, a global/total error rate is simply obtained as the sum of false positive rate and false negative rate [Sha99].

When analyzing acoustic speech, segment boundaries may appear between any pair of adjacent feature frames. This contrasts to symbolic speech where boundary markers are only permissible between phone, syllable or word symbols, which makes the number of possible segment boundaries to be a magnitude smaller than for acoustic speech processing. Consequently, above mentioned metrics should be applied carefully. It is necessary to replace the hard boundaries of a reference annotation with sufficiently softened markers. Most commonly, detected segmentation markers are treated as correct, if they appear within a certain time-frame around a true marker [Tol04]. The derived metric is than chosen to be the percentage of boundary markers with errors smaller than the chosen tolerance window.

7.1.3 Kappa

Precision and recall do not take into account that any classifier that assigns patterns to clusters will have a certain ratio of chance correctness depending on the pattern distribution and the number of clusters. This is taken into account by the κ statistic. Basically, it allows to compare two classifiers that assign N pattern to K distinct classes. It measures the agreement between these classifiers and is calculated by

$$\kappa = \frac{P_A - P_E}{1 - P_E} \tag{7.7}$$

Hereby, P_A and P_E denote the probabilities for relative observed agreement and chance agreement respectively. Commonly, κ is computed to compare a ground truth clustering against an actual classification of a pattern set. Because of its definition, κ is distributed in [0...1], whereby higher values indicate a better agreement between both classifiers.

7.1.4 Segmentation vs. clustering

To overcome the problem of discrete markers in a quasi-continuous feature stream, any segmentation can be regarded as a clustering problem. Whereas a corpus annotation provides just one possible ground-truth clustering, every other parsing into somehow labeled segments defines an alternative clustering. A powerful information theoretic means to compare any pair of clusterings is *Variation of Information* $\mathcal{VI}$ proposed in [Mei02]. $\mathcal{VI}$ measures the amount of information that is lost or gained when transforming a clustering $\mathcal{C}$ into a clustering $\mathcal{C}'$. It is computed as (cf. sec. 3.3)

$$\mathcal{VI}(\mathcal{C},\mathcal{C}') = \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C},\mathcal{C}') \tag{7.8}$$

Interestingly, $\mathcal{VI}$ satisfies all conditions of a metric. Furthermore, it allows to compare clusterings obtained with different numbers of classifiers. This especially applies to unsupervised learning, where the number of clusters emerges naturally from the data.

7.1.5 Statistical learning quality

Metrics to score n-gram models are supposed to reflect the amount of statistical constraints captured from the training data. Clearly, this depends on the amount of structure in the data itself. For instance English can be considerably better modeled using n-gram models compared to German or French, because of its relatively linear utterance structure and missing flectional forms. In other contexts like digit recognition n-gram model are hardly applicable at all.

A popular metric to assess the amount of transition statistics encoded in an n-gram model L is *perplexity*. It is calculated with respect to a test-sequence $\mathbf{w} = w_1, ...w_T$ by

$$PP(\mathbf{w}) = \frac{1}{\sqrt[|\mathbf{w}|]{\mathrm{L}(\mathbf{w})}} = \mathrm{L}(\mathbf{w})^{-\frac{1}{T}} \tag{7.9}$$

Briefly, perplexity measures the average branching factor, which is the average number of successor symbols. Lower values PP usually correlate with better system performance in ASR systems, because of the better prediction of successor elements. When there are no structural constraints in the transition structure of a time-series, $\mathrm{PP}(\mathbf{w})$ equals the number of items in the symbol inventory. In ASR applications PP typically varies between 10 (for digits) and 1000 (for generic n-gram word models of English). Because statistical modeling takes place on different levels of granularity in our framework, we denote perplexities for word, syllable and phone layer as PP_W, PP_S and PP_P respectively.

7.2 Corpora

The model proposed in this thesis is designed to capture the structure of speech on different levels of granularity. Thus, we employed different special-purpose corpora to assess the performance of the different sub-systems. The common theme of these corpora was to reflect some properties of infant-directed speech. To ease evaluation and to reveal possible limitations of our approach, input speech would optimally obey defined statistic regularities. This renders common speech corpora to be not suited for the evaluation of our model. This is because neither their structure is known in terms of phones, syllables and words, nor is it possible to adjust their frequency and transitional pattern distributions according to different evaluation hypotheses.

7.2.1 Phones

Phones are hardly observable without a surrounding speech context. Thus, we recorded a small dedicated phoneme-corpus. This was driven by the need to evaluate the proposed phone clustering in a systematic way. We selected those phonemes for recording which also matched our phone definition given in chapter 2. The set of recorded phonemes comprised most fricatives, nasals, semi vowels, slides and vowels as defined in the TIMIT phoneme inventory. Stops were not included. Overall 22 phones were recorded as stationary sounds without any speech context 20 times each.

7.2.2 Monosyllabic words

According to studies of Greenberg [Gre98] the 30 most frequent words in the large English SWITCHBOARD speech corpus are monosyllabic. Furthermore, the most 100 frequent words include only 10 poly-syllabic words. Stated another way, 81% of all used words are mono-syllabic [Gre98]. According to the findings summarized in chapter 2, CDS can be assumed to contain an even a larger proportion of mono-syllabic words.

Hence, we recorded a speech corpus *MonSyl* containing 30 monosyllabic words. This corpus was recorded by a single speaker under low-noise conditions using a headset microphone. Each syllable was recorded 100 times without any speech context to ease automatic energy based segment extraction. The resulting data-set was annotated automatically on syllable level.

7.2.3 Semi-synthetic speech

To investigate speech acquisition principles that rely on continuous speech, we have developed a semi-synthetic speech generator/tutor SYTU. It can be defined as a generative probabilistic syllable grammar with syllable-node dependent output distributions that emit acoustic speech segments. Thereby, emitted syllable speech segments were sampled randomly from the recorded syllable corpus *MonSyl*. Generated symbolic syllable sequences became converted to give acoustic speech utterances by concatenating these speech snippets. Segment annotation files were generated automatically during this generative process.

In contrast to text-to-speech systems this gives a diminished speech quality. This is mainly due to the trivial speech segment concatenation method implemented for our model, and the lack of any control over intonation contours. However, SYTU comes along with the important property that different realizations of the same utterance differ on the acoustic level, which is a prerequisite to evaluate the properties of the proposed model for acoustic speech structure acquisition.

7.2.4 Discrete speech

To gather first insights into the process properties of the proposed architecture, and to obtain a base line of performance, we employed generative probabilistic grammars to create symbolic phone as well as syllable input for our system. To ensure comparability, all 30 words contained in *MonSyl* (see below) were embedded into a flat rule grammar *PSSG*. To setup a probabilistic phone grammar *PPSG* we replaced the word symbols with phoneme symbols as defined by the TIMIT corpus.

In case of a flat probabilistic grammar, the number of syllables contained in the generated utterances was limited by a Gaussian length model on the number of syllables. Furthermore, we added for both models the possibility to add arbitrary amounts of phone/syllable substitution noise. This was necessary to take the imperfect nature of syllable and phone recognition into account. For simplicity, replacement probabilities were not chosen to depend on the acoustic class of the respective symbol, but were uniformly distributed over the complete speech unit space.

7.2.5 Child-directed read speech

Neither the semi-synthetic speech tutor SYTU nor the isolated words corpus *MonSyl* are suited to investigate the development of childlike speech processing abilities to full extent. This would require their complexity and structure to be magnitudes more diverse. Even under the (unlikely) assumption, that the model proposed in this thesis actually corresponds to the way in which human speech structure acquisition is being organized, we do not consider a purely acoustic speech corpus to be sufficiently rich given the multitude of speech and non-speech cues that are available to the young learners. Whereas promising first steps to record such a corpus by eavesdropping a complete environment of an infant are currently underway [Roy06], a mapping to these recordings to the actual sensory input signal of the young learner remains an open question and subject to ongoing research [Yu08]. Furthermore, such a corpora would need to be exhaustively annotated in order to assess the quality of the emerging speech representation.

Although some efforts lead to corpora like CHILDES [Mac95] or MOTIONESE [Roh04], we consider these to be only partially suited for our purpose because of two reasons. First, even if such corpora contain child-directed speech, they usually encompass a large set of speakers. This clearly contrasts with the experience of infants that mainly interact with a small set caregivers. Second, and even more important, we are not aware of corpora [1] that provide a sufficiently large speech sample that also encompasses the structural changes of CDS during development of the infant.

To come closest to a realistic setting of speech acquisition, we use large amounts of read speech because it tends to be more carefully realized compared to spontaneous speech. We have chosen the recordings of a collection of poems for children [Var06]. These poems have been read by a single female speaker and have been published as part of the LIBRIVOX-project [Var09]. Even if the source texts are freely availably, the recordings are not annotated. The total length of them is over 8 hours.

7.3 Phones

There are two major aspects to be evaluated concerning the phone layer as described in section 6.3. First, the proposed learning scheme has to be evaluated whether is suitable to reveal a phone representation in an unsupervised manner. Next, the properties of the incrementally bootstrapped phone n-gram model have to be investigated.

[1] With the notable exception of [Roy06], which would perfectly fit our needs. However, we were told that this corpus is not available for research outside the initiator's institute.

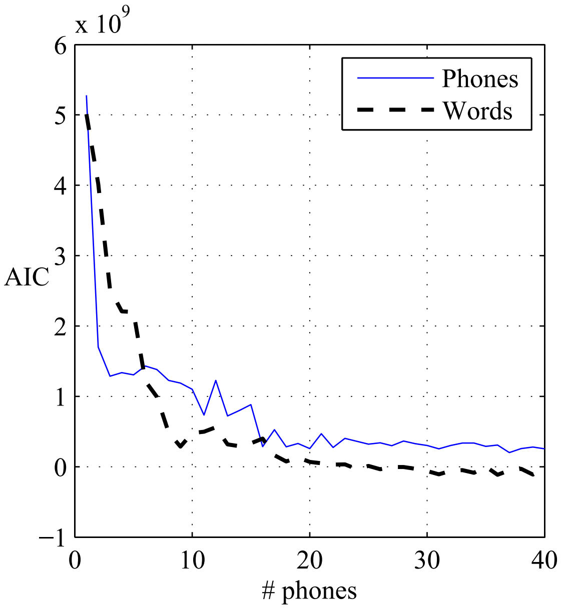


Figure 7.1: AIC as function of the number of phone models. The dashed and the solid line denote the word and phone test data case respectively. Under both conditions the elbow function saturates at around 20 to 25 phone models.

7.3.1 Classification

First we computed a base-line for phone classification by training a supervised phone model $\mathcal{M}_P^S$ on a subset of the phone recordings corpus described in section 7.2.1. Applied to a corresponding test set the detection rate was 100%, which was not surprising as the task was to match 22 well trained phone-models against rather stationary test patterns.

Next, we bootstrapped a phone representation using the unsupervised phone clustering method proposed in section 6.3. Thereby, the number of phone models was kept fixed to the number of corpus phones (22). By applying the model labeling technique described in section 7.1.1 we obtained a matching rate of 28.8%.

As described in section 6.3.1 the optimal number of phones can be determined using the AIC as goodness of fit measure. Figure 7.1 shows the elbow-functions with the number of phones as independent variable in case of unsupervised learning. We evaluated two different data conditions to assess the influence of co-articulation effects on the discriminative function of the acquired phone models: Portions of the phone corpus 7.2.1 were used to bootstrap $\mathcal{M}_P^P$, and a subset of the monosyllabic word corpus 7.2.2 was used to estimate $\mathcal{M}_P^W$. For AIC calculation we used a distinct subset of $MonSyl$. As AIC saturates at around 20 to 25 phones under both conditions, a phone pool with a size of around 25 gives a reasonable trade-off between model complexity and computational decoding demands.

7.3.2 Clustering

In the previous section we focused on isolated phone classification. But this is a rather artificial scenario, as phones rarely appear without any speech context as in our phone classification experiment. Thus, we evaluated the segmentation performance of the phone decoder. The test set was chosen to be a subset of $MonSyl$ comprising 500 syllable segments.

Table 7.1 summarizes the results of the experiment. Three different phone representations were investigated: $\mathcal{M}_P^P$ and $\mathcal{M}_P^W$ as in the last section. Additionally we included the decoding performance when using $\mathcal{M}_P^S$ as base-line. Table 7.1 reports variation of information $\mathcal{VI}$ with respect to the segmentation obtained from $\mathcal{M}_P^S$.

	$\mathcal{M}_P^S$	$\mathcal{M}_P^P$	$\mathcal{M}_P^W$
$\mathcal{VI}$	8.8E-16	3.964	4.95

Table 7.1: Comparison of phone segmentations obtained by using differently trained phone models. The statistics are computed against a reference segmentation as provided by $\mathcal{M}_P^S$. The first column only indicates a perfect match which sums not exactly to 0 because of the sampling process necessary for $\mathcal{VI}$ calculation.

7.3.3 Phone-distributed word models

To further assess whether a phone representation can be derived in a purely bottom-up manner, it is necessary to investigate the quality of the bootstrapped phone representation. We followed the evaluation scheme proposed in [Mar07]: for a set of isolated annotated word utterances, the decoded phone sequences were labeled and retained.

For testing, decoded phone-sequences were compared to all phone sequences attached to the database words. This was realized by computing the edit-distance between each test sequence and all collected training sequences. The word associated to the best matching training sequence was considered as recognition result. In contrast to stationary sounds as used in [Mar07] we assessed the model quality in a more sophisticated scheme by testing monosyllabic words and not only letters as reference-units. For evaluation we used the 30 monosyllabic words contained $MonSyl$. Training and test sets contained 50 instances each of each word.

The results are summarized in table 7.2. Figure 7.3.3 depicts the corresponding classification confusion matrix in case of using $\mathcal{M}_P^W$. Because of co-articulation effects between subsequent phones, distributed word models show only little discriminative power when being applied to a word classification task.

Used Model	$\mathcal{M}_P^S$	$\mathcal{M}_P^P$	$\mathcal{M}_P^W$
(1 - WER)	26%	13%	29%

Table 7.2: Isolated word recognition rates when using a phone-sequence model for classification. As expected the co-articulation effects between subsequent phones render the distributed word models to be not discriminative in a word classification task. This also holds for supervised trained phones model $\mathcal{M}_P^S$ which excludes an ineffective phone clustering approach as possible cause for the high classification error. These results confirm our initial assumption that syllables (and thus words) need to be modeled as distinct acoustic models and not as concatenated phone models.

7.3.4 Phone language model

As described in section 6.3, a probabilistic phone-grammar L_P is learned incrementally while processing the speech input. To assess its quality we tracked perplexity during the bootstrapping process (cf. section 7.1.5), which was updated with a frequency of 10 training utterances.

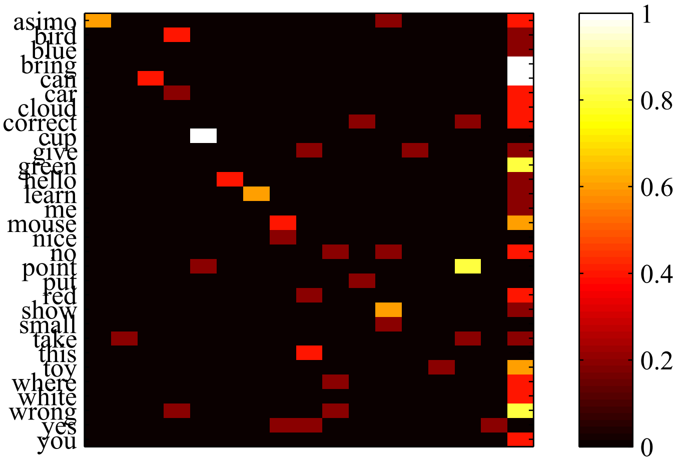
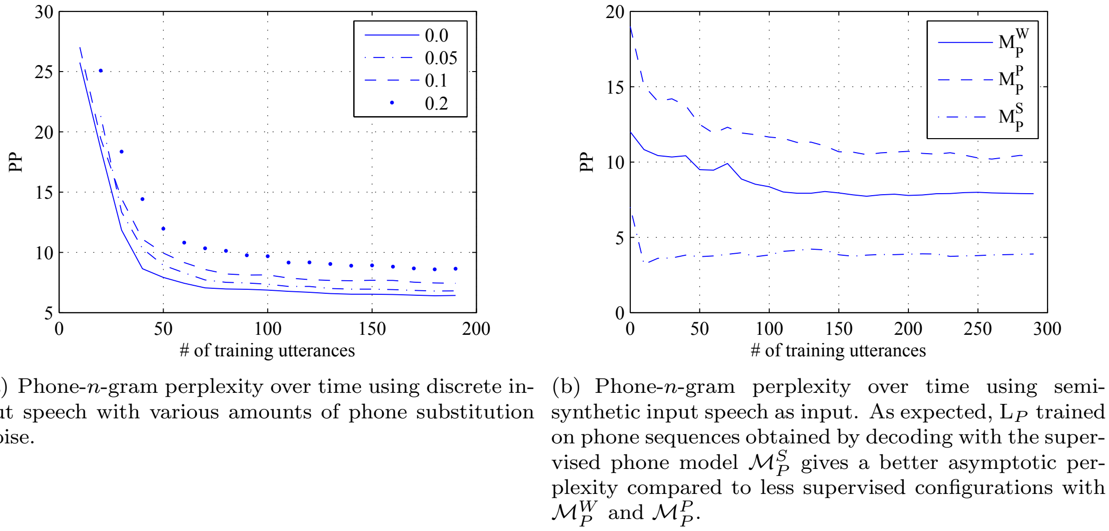


Figure 7.2: Word classification confusion when using distributed phone-sequence models with $\mathcal{M}_P^W$ as phone representation.

Figure 7.3(a) depicts perplexity functions obtained by processing discrete phone sequences sampled from *PPSG* with various amounts of phone substitution noise as described in section 7.2.4. Next, perplexity was tracked for each of the different phone representations $\mathcal{M}_P^W$, $\mathcal{M}_P^S$ and $\mathcal{M}_P^P$ on speech utterances generated with SyTu. The resulting perplexity functions are depicted in 7.3(b). Finally, phone language model perplexity was tracked for unconstrained read speech sampled from Librivox as input. The used phone model $\mathcal{M}_P^L$ was trained with the unsupervised phone learning mechanism prior to the learning of L_P on a distinct sub-set of Librivox. Figure 7.4 depicts the resulting perplexity function.

For all types of phones and all types of input speech, perplexity converges as L_P stabilizes. Because of the obtained asymptotic perplexity values, L_P seems to successfully capture phone transition constraints.



(a) Phone-n-gram perplexity over time using discrete input speech with various amounts of phone substitution noise.

(b) Phone-n-gram perplexity over time using semi-synthetic input speech as input. As expected, L_P trained on phone sequences obtained by decoding with the supervised phone model $\mathcal{M}_P^S$ gives a better asymptotic perplexity compared to less supervised configurations with $\mathcal{M}_P^W$ and $\mathcal{M}_P^P$.

Figure 7.3: Phone perplexity for different types of input speech.

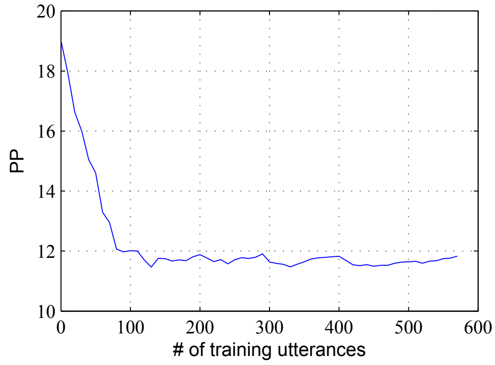


Figure 7.4: Phone-n-gram perplexity over time for unconstrained read speech as input.

7.3.5 Phonotactic model

To evaluate how speech can be framed into syllable segments using the proposed phonotactic parser, it is necessary to investigate the properties of the phonotactic model P_{SB}. We focus on two major aspects:

1. What are the properties of the bootstrapping process?
2. What are properties of the converged phonotactic model $P_{SB}(t \to \infty)$? Does $P_{SB}(t \to \infty)$ capture a sufficiently rich model of the tutoring language's phonotactics to build a syllable parser/segmenter?

Process properties

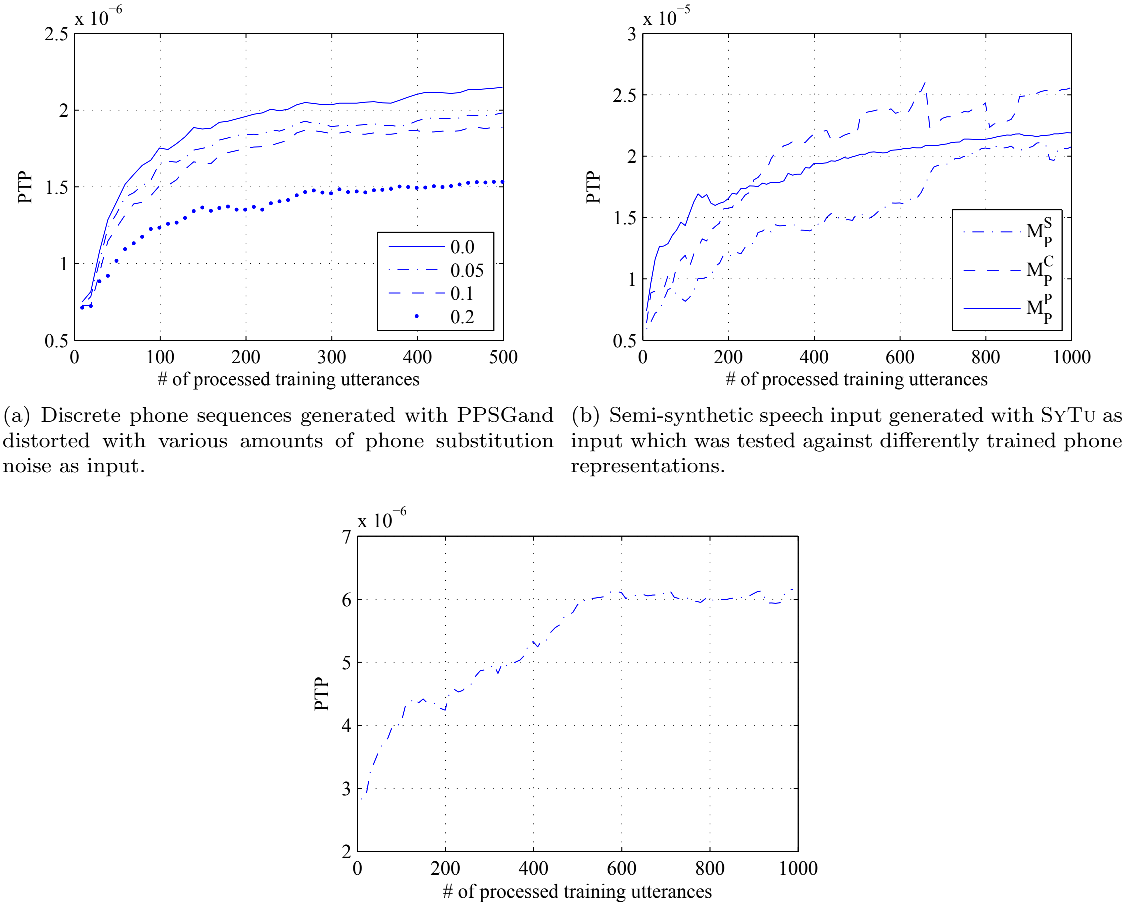
We follow the ideas of [Chr98, p. 20] to rate the performance of the P_{SB} according to its predictive power. This can be assessed, by investigating to which extent P_{SB} is suitable to predict syllable boundaries. For this purpose we define *phonotactic perplexity* $\mathcal{PTP}$ as average phonotactic probability over a set of utterance boundary phone sequences $\mathbf{u} = u_1, u_2, ..., u_N$ as follows:

$$\mathcal{PTP}(\mathbf{u}, P_{SB}) = \frac{1}{2N}(\sum_{i=1}^{N} P_{SI}(\oplus B_I(u_i)) + \sum_{i=1}^{N} P_{SF}(B_F(u_i)\oplus)) \tag{7.10}$$

Thereby $B(u)$ denotes the boundary function that extracts the boundary part of a phonified utterance u. For P_{SI} and P_{SF} the boundary functions $B_I(u)$ and $B_F(u)$ are used respectively.

We updated $\mathcal{PTP}$ with a frequency of 5 utterances to assess how the predictive power of P_{SB} evolves over time. As for perplexity, three different types of input speech were compared. First, we investigated discrete phone sequences sampled from $PPSG$ with various amounts of phone substitution noise. This was motivated by the need to obtain a base-line performance in case of a (semi)perfect phone decoder. Second, semi-synthetic speech generated with SyTu was evaluated with different phone representations. Finally, we used portions of Librivox as input along with $\mathcal{M}_P^L$ as phone representation while tracking $\mathcal{PTP}$.

Figure 7.3.5 depicts how $\mathcal{PTP}$ evolved in the different experiments. In all configurations, it first increases monotonously and converges against a speech-input-dependent saturation level. For $\mathcal{PTP}$ calculation we used a distinct subset of size $|\mathbf{u}| = 1000$ sampled from each evaluation corpus



(a) Discrete phone sequences generated with PPSGand distorted with various amounts of phone substitution noise as input.

(b) Semi-synthetic speech input generated with SyTu as input which was tested against differently trained phone representations.

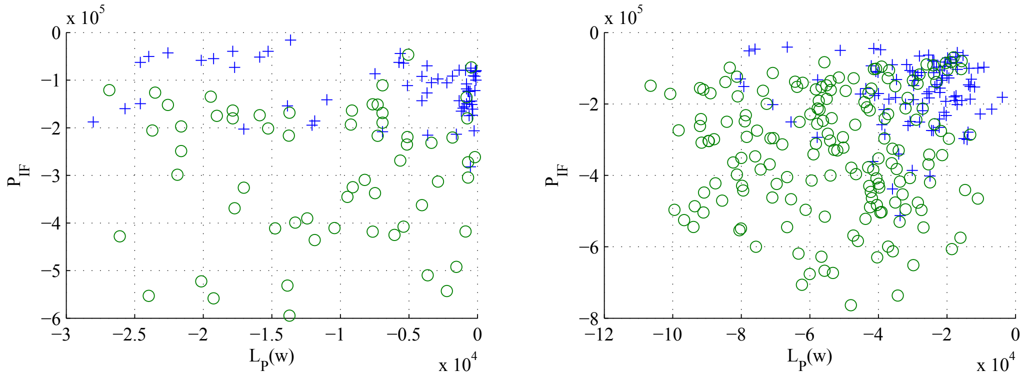
(c) Child-directed read speech sampled from Librivox as input. The phone model P^U_{SB} was used for decoding.

Figure 7.5: Tracking of phonotactic perplexity for different types of input speech under various evaluation conditions.

respectively.

This last experiment shows that the representation converges against a stable state under a wide variety of input and evaluation conditions. However, to test whether the captured constraints about the tutoring language are actually sufficient for syllabic parsing, the next evaluation was designed to validate that the model clearly discriminates between the tutoring language and a random test language. For this purpose we randomly sampled phone sequences with the same phrase length statistics as used for *PPSG* from the phone symbol inventory $\mathcal{M}_P$ as acquired by the system. Next, we scored each of these sequences as well as a set of phonified tutoring language utterances with respect to (a) hone n-gram probability and (b) average phonotactic probability. Thereby, the latter was defined to be the mean of initial and final phonotactic probability and was calculated by $P_{IF}(w) = \frac{P_{SI}(w)+P_{SF}(w)}{2}$.

Figure 7.3.5 depicts the results. The x position shows $\mathrm{L}_P(w_i)$ and the y position denotes $P_{IF}(w_i)$, $i = 1, \ldots, 200$. For each type for input speech, the matching test sample gives significantly higher scores compared to the random language sample. The former forms a clear cluster in the upper right corner, whereas the latter is uniformly distributed without any clear structuring.



(a) System was trained with 2000 phrases sampled from SyTu. The used phone-model was $\mathcal{M}_P^W$.

(b) System was trained with 2000 phrases sampled from Librivox. The used phone-model was $\mathcal{M}_P^L$.

Figure 7.6: Language identification using phonotactic and statistical cues. After familiarizing the system with a tutoring language, we exposed it to 2000 utterances of the tutoring language itself and a random test language. Each utterance was scored individually with P_{IF} and L_P and visualized as scatter dot. The tutoring and the random test language are indicated by crosses and circles respectively. For both evaluation conditions, 200 samples of the target and the random language were used for testing. Each point in the figures represents a single test utterance w. The position of the utterance in the scatter plot is defined by the averaged phonotactic probability $P_{IF}(w)$ and the utterance phone-language model probability $\mathrm{L}_P(w)$. The overlap between both distributions follows from the definition of the random language which models all possible symbol sequences with equal probability.

7.3.6 Syllabic parsing

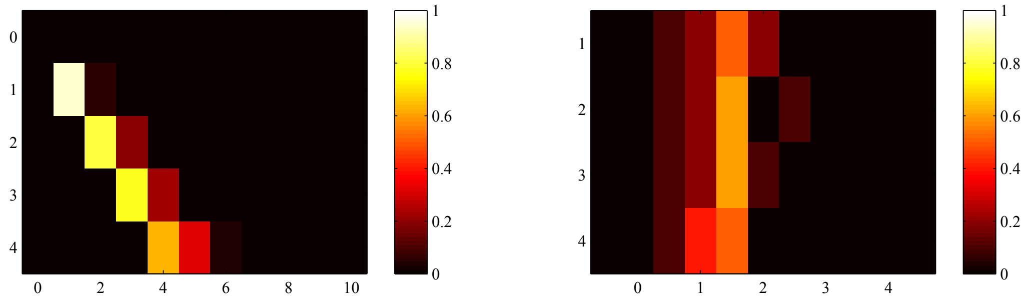
Next, we evaluated the actual framing of continuous speech into syllabic segments by means of the proposed phonotactic parser. First, to obtain a baseline of performance we generated phone sequence utterances with $PPSG$ comprising between 1 and 4 syllables. Prior to testing, the system was trained with 2000 of these utterances to bootstrap the phonotactic model. For testing a distinct sample that was processed by the syllabic parser as described in 6.3.2 to reveal the number of syllables contained in each test sequence.

To evaluate phonotactic parsing we faced a dilemma. On the one hand, it would have been clearly beneficial to profit from the large amount of statistics captured on the Librivox corpus 7.2.5. However, without any annotation there is no way to automatically assess the quality of the parsing results. Hence, we were only able to test phonotactic parsing using synthetic speech utterances generated by SyTu as input. Those became converted into phones sequences (cf. 6.3), which were subsequently processed by the parser.

Figure 7.3.6 depicts the results as confusion matrix. The system is able to reveal the number of syllables per utterance with high reliability in the case of discrete phone input. In case of continuous speech the performance is reduced, as the system often overestimates the number of syllables.

Context size

Next we evaluated the context size of the P_{SI} and P_{SF}. This needs to be a trade-off between structural constraints that span over different phones, and the unknown phone complexity of the syllabic structure under question. As the latter may vary from single up to many phones, a too wide context is likely to include more than one syllable, which would result into a poor probabilistic



(a) Prior to testing the system was trained with 2000 phrases sampled from *PPSG*.

(b) Prior to testing the system was trained with 2000 phrases sampled from Librivox. The used phone-model was $\mathcal{M}_P^L$.

Figure 7.7: Phonotactic parsing of different types of input speech. For both types of input speech, the number of syllables per phrase was varied between 1 and 4.

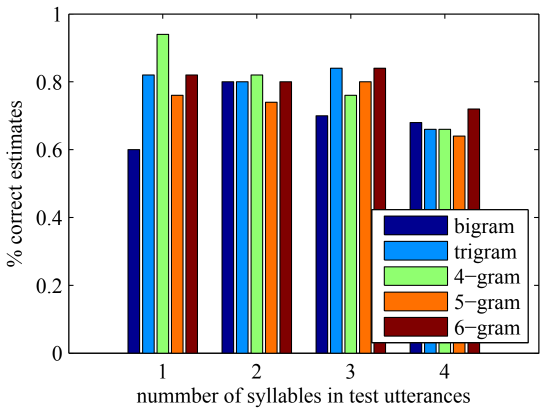


Figure 7.8: Detection rates for different context sizes and different syllables per test utterance.

encoding of the language phonotactics.

To investigate the effect of the n-gram context size, we compared the ratios of successful parsings for different context sizes varying between 2 and 6 on speech phrases consisting of 1 to 4 syllables (with different phonological complexities). Figure 7.8 summarizes the results. As described in 6.3.2 the syllable boundary symbol needs to be attached as final symbol in the n-gram context. Hence, the actual phone-context is reduced to $n - 1$. For instance, a bi-gram model corresponds to a phonotactic uni-gram model, which takes only the boundary phone symbol into account. The system was evaluated with discrete phone utterances generated with *PPSG* and acoustic phrases as provided by SyTu. Both were constrained to generate utterances with 1 to 4 syllables. To match the properties of phone sequences obtained from acoustic speech, discrete test-sequences were distorted with a phone-substitution noise of 10%. The number of training utterances to bootstrap the phonotactic model in each context-configuration was chosen to be 2000. As shown in figure 7.8, a context size of 4 slightly outperforms the other configurations in the important case of a single syllable. However, a clear trend is not prominent, as the trainability of P_{SB} decreases for larger context sizes.

7.4 Syllables

Comparable to phones, we consider the properties of the bootstrapping process, and the discriminative function of the resulting syllable representation to be the most relevant aspects for evaluation. In addition, the effects of the different acquisition principles, as well as the performance of the syllable spotter need to be investigated.

Most our experiments require some kind of syllable annotation. Thus, we usually run the experiments against portions of *MonSyl* or synthetic utterances generated by SyTu. Where feasible, we furthermore validate the model on unconstrained speech sampled from Librivox. As for phones, we fall back to use discrete syllable sequences (if necessary) to obtain performance base lines.

7.4.1 Model initialization

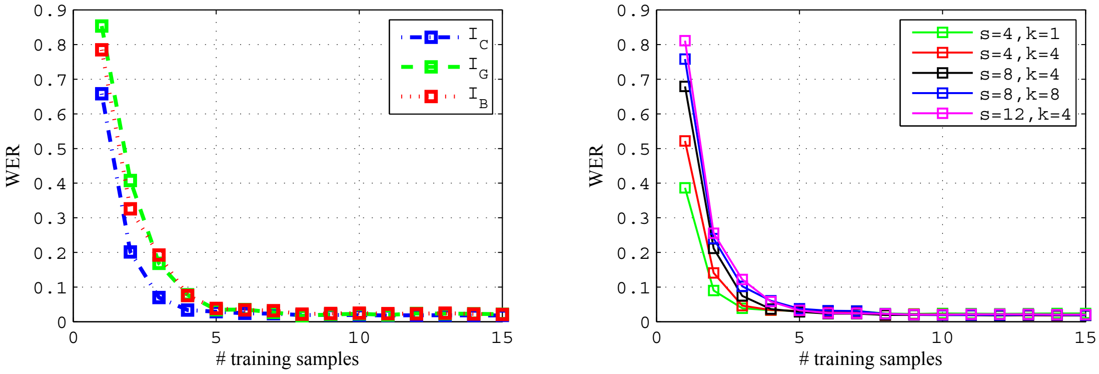
Model initialization plays a critical role in HMM learning as pointed out in section 3.4.1. Thus, we compare different settings to reveal the optimal procedure. The phone-concatenation method introduced in section 6.4.3 for syllable model initialization is referred to as I_C.

Another method for model initialization I_G, is to adapt a new syllable model from a generic syllable model λ_S^G estimated on an arbitrary (but distinct) speech sample (of $MonSyl$). Because it can not be assumed that all those initial training segments contain the same word, this model has to be considered as a rather generic syllable structure model and not as a model of a specific syllable.

A natural choice for model initialization within an incremental clustering framework like the one proposed in alg. 6.9 is to derive a new model from the best matching model λ_S^*. Thus, we define the model initialization method I_B, which clones and adapts λ_S^* to create a new model. In terms of clustering, the adaption of such a cloned cluster with the segment X causes shift of the model parameters into the direction of the underlying syllable structure. To make I_B to fit into the used offline evaluation scheme of this experiment, we refer the pool of existing model as the models estimated prior to the syllable in question. As the first syllable has no predecessor λ_S^G is included as possible predecessor for initialization but not for classification.

For the actual adaption of the initialized models we used in all cases the EM scheme described in section 6.4.3. Syllable models were trained supervised for 30 (mostly) monosyllabic short words contained in $MonSyl$ using the different initialization methods. To take the amount of training data into account we varied the number of training segments between 1 and 15. The number of test segments per model was 85. Figure 7.9(a) depicts the WER-curves as obtained from the experiment. The proposed concatenative initialization method I_C outperforms the other methods especially for the important case of very few training samples. For more training samples the different methods perform equally well.

Next, to assess the influence of the syllable model topology we tested different configurations of the number of HMM states and the number of component densities for each state. I_C was used as initialization method and the training and test conditions were kept as in the last experiment. The results are depicted in figure 7.9(b). Configurations with fewer model parameters tend to outperform more detailed models. This is mainly because of the better trainability of less complex models in case of few training data.



(a) WER as function of the number of training segments evaluated for different types of model initialization.

(b) WER as function of different topology configurations in terms of HMM-states s and component densities k.

Figure 7.9: WER-plots for different HMM-topologies and initialization methods.

7.4.2 Clustering process properties

To investigate the properties of the proposed syllable bootstrapping scheme 6.9, we processed sub-sets of $MonSyl$. The used phone-model was $\mathcal{M}_P^W$ and the initialization for new syllable unit models method was chosen to be I_C. The properties of the emerging syllable representation were sampled with a fixed frequency of 10 syllables. To assess the process properties we logged the following statistics:

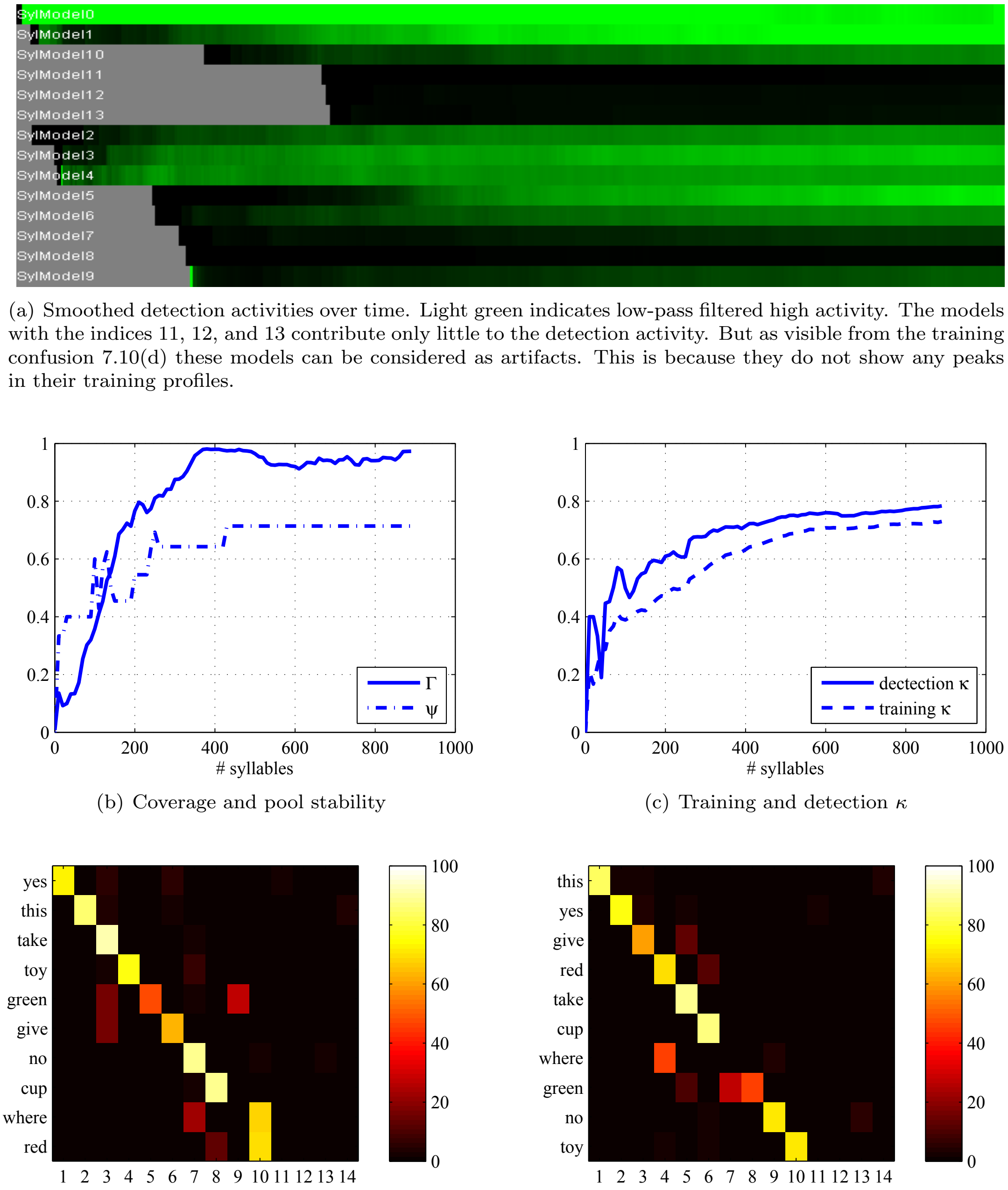
1. Coverage Γ and pool stability ψ as these quantities define basis of the proposed regulatory framework.

2. Model activity traces that show frame-normalized detection activities smoothed over a $30s$ sliding window. This we found to be an appropriate means to visualize how the distribution of model activity evolves over time. Especially it allows to identify 'dead' models that do not contribute to the segmentation of the speech signal. Furthermore, this plot allows to identify the creation of new syllables models.

3. Training and detection κ calculated from the confusion matrices of training and classification. Both are built using the corpus annotation and the reordering scheme 7.2. In contrast to the two former statistics, these κ-values are not available to the system during clustering and are only logged for performance tracking.

Figure 7.4.2 depicts the results obtained while processing a sample of the 10 words subset $MonSyl_{10}$ comprising 900 syllables instances. The unsupervised clustering allows to identify all syllables with good accuracy. The progression of the regulatory measures 7.10(b) shows that the syllable representation converges against a stabilized state. Speech coverage Γ is initially dominated by filler segments but converges against 100% as the syllable representation stabilizes. A similar progression is observed for training κ. Detection κ corresponds to the average model selectivity during syllable spotting, and is also converging.

As visible from the activity traces 7.10(a), some syllable models do not contribute to the segmentation activity. However, by taking the training confusion 7.10(d) into account, this becomes reasonable, as the inactive models do not clearly correspond to any of the input syllables. Beside these 'dead' models (that are possible candidates for an additional pruning process), training and detection confusion show that the clustering process reveals highly discriminative and specifically



(a) Smoothed detection activities over time. Light green indicates low-pass filtered high activity. The models with the indices 11, 12, and 13 contribute only little to the detection activity. But as visible from the training confusion 7.10(d) these models can be considered as artifacts. This is because they do not show any peaks in their training profiles.

(b) Coverage and pool stability

(c) Training and detection κ

(d) Training confusion

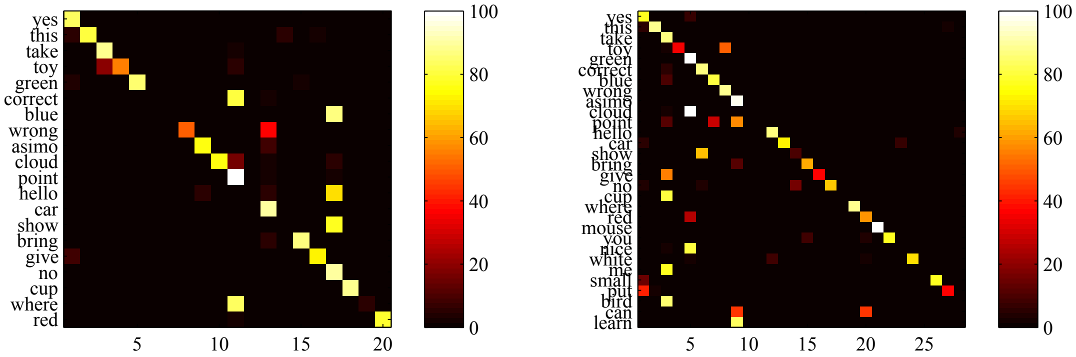
(e) Detection confusion

Figure 7.10: Properties of the syllable clustering process when processing $MonSyl_{10}$ as input.

trained syllable models in almost all cases. Existing mis-assignments in training and detection are often due to phonological similarities between test and evaluation patterns.

Effect of parameterization

Next, we aimed to assess whether regulation takes place as expected, or whether the system parameterization accounts for the results obtained in the last experiment. Two larger subsets $MonSyl_{20}$, $MonSyl_{30}$ of $MonSyl$ comprising 20 and 30 randomly ordered monosyllabic words respectively are used as input to the system. The final training confusion matrices are shown in figure 7.11.



(a) Training confusion for $MonSyl_{20}$.

(b) Training confusion for $MonSyl_{30}$.

Figure 7.11: Syllable training confusion matrices for corpora of different size. The number of elements in the processed corpora correlates well with the number of syllables models emerged from the unsupervised incremental learning process. The parameterization of the system was the same as used for fig. 7.4.2

Using the same parameterization of the different modulatory processes as for the previous experiment 7.4.2, the emerging representation reflects the complexity of the input speech, which has been proposed in [Cer08] to be a valid metric for morpheme acquisition. For $MonSyl_{10}$, $MonSyl_{20}$ and $MonSyl_{3}0$ the clustering algorithm reveals 14, 20 and 28 syllables models respectively, which we consider to be a surprisingly good result for an unsupervised learning approach given the complex structure of the input space. This supports and validates the proposed regulatory framework for syllable representation learning.

Stability vs. Plasticity

To assess the stability of the learned syllable representation we applied a corpus switching scheme. The monosyllabic word corpus $MonSyl$ was split into 3 syllable-disjunct subsets $MonSyl_n, n \in A, B, C$, each comprising 1000 instances of 10 syllables. The system was instrumented to switch between these disjunct subsets after they had been completely processed, which totally exchanged the syllable input the system was perceiving.

Figure 7.12 summarizes the results. It plots the best matching model likelihood $l(X|\lambda_S^*)$ calculated for each input segment X, and the syllable pool-size $|\mathcal{M}_S|$ against the number of processed syllables. The changes of the input signal are clearly visible in both figures. The first change (t=1000) reflects in a major decrease in $l(X|\lambda_S^*)$, which is not as prominent for the second change because of the increased inherent acoustic variability of $\mathcal{M}_S$ at $t = 2000$. Both changes match very well with the number of models, which converge against the actual number of distinct syllable entities in the corpus. As the latter is not observable to the system, the proposed clustering approach can be considered to be functional even under changing input conditions.

Next, we proved that the emerging syllable representation is not only plastic enough to cope with changing input, but keeps its knowledge about previously learned patterns. We tested the discriminative function with respect to $MonSyl_A$ of the $\mathcal{M}_S(t = 3000)$ after all three subsets have been processed. By using the maximum-trace permutation scheme proposed in section 7.1.1, we labeled all models in $\mathcal{M}_S(t = 3000)$ and computed the WER with respect to a disjunct test $MonSyl_A^T$ ($|MonSyl_A^T| = 1000$) set comprising only examples of the syllables in $MonSyl_A$. The obtained WER was 21.3%. Because the labeling algorithm disregards other models for a particular

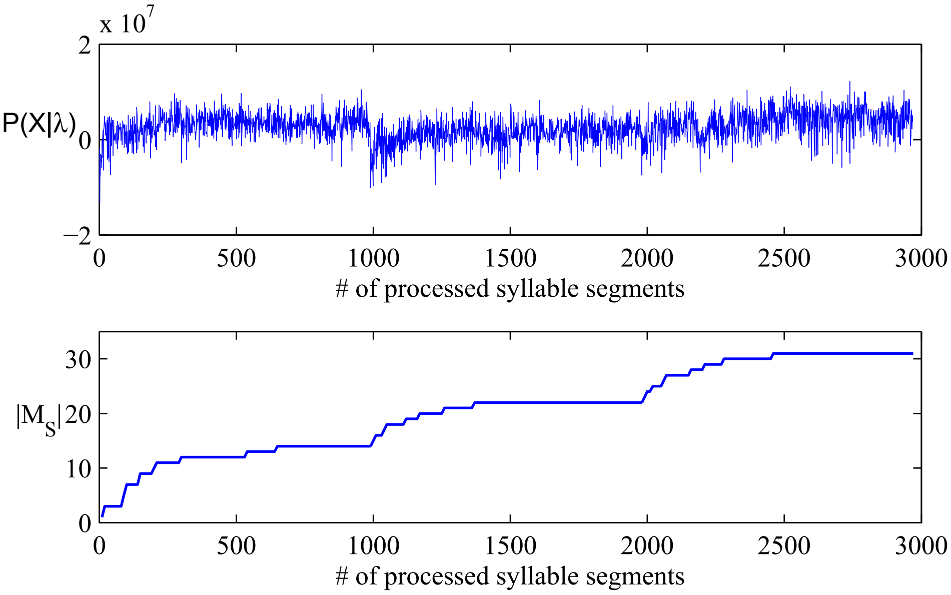


Figure 7.12: Effect of switching the input between distinct sub-sets of *MonSyl* comprising 10 syllables each. The system configuration was kept unchanged and the system was not being made aware of the switch beside of the changing input signal. The system realizes the drop-back in matching performance, and initiates the creation of new syllable clusters. For all three subsets the system converges (approximately) against the correct number of syllables that have been presented since the *start* of the clustering process. Hence, the proposed clustering approach manages to keep plasticity while maintaining the stability of already acquired syllables models.

syllable than the trace-maximizing one (in case of over-representation), we consider this to be a very good result given the complexity of the task.

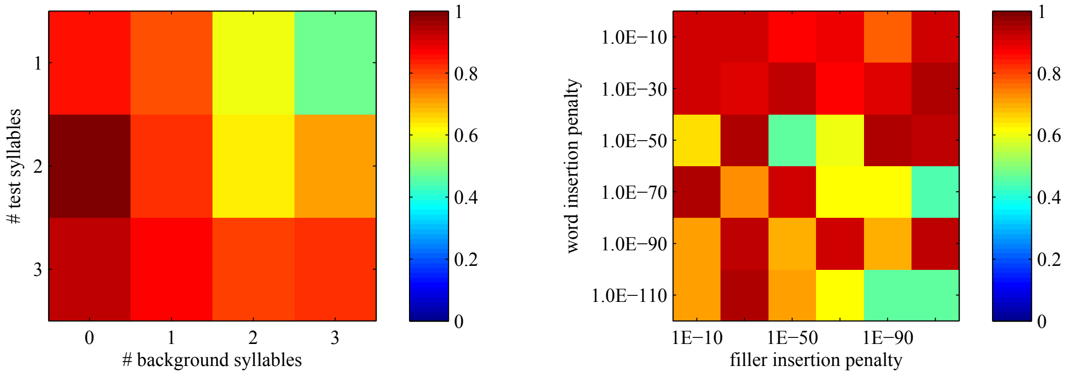
7.4.3 Spotting performance

As the syllable detection module is implemented as a keyword spotter, we tested the spotting-performance using semi-synthetic speech generated by the synthetic speech tutor SyTu. For evaluation we split the syllables contained in *MonSyl* into a test S_T and a background-set S_B containing 20 and 10 syllables respectively. Using a flat uni-gram model we generated utterances comprising N test and M background syllables. Figure 7.13(a) visualizes the detection results in terms of F_1 for utterances of different complexity. The order of test and background words was hereby sampled randomly. The obtained spotting performance is similar to what has been reported for state of the art keyword spotting approaches as outlined in section 3.6.3.

As syllable spotting performance relies on the choice of insertion penalties, we varied syllable insertion penalty E_W and filler insertion E_F independently. The obtained results in terms of F_1 are depicted in figure 7.13(b). The number of test and background syllables per utterance was randomly sampled from $\{1, 2, 3\}$ and $\{0, 1, 2, 3\}$ respectively. The performance increases with lower word insertion probabilities. For filler insertion penalties no clear trend was observable, which we assume to be caused by an (non-identified) systematic evaluation problem.

7.4.4 Subtraction learning

As we consider the principle of subtraction to be a core principle of speech acquisition (cf. 2.2.3), it is crucial to assess our system's abilities to extract segments from continuous utterances (cf.



(a) Spotting performance in terms of F_1 for utterances comprising different numbers of test and background syllables. The system was parameterized with $E_W = 1 \cdot 10^{-75}$ and $E_F = 1 \cdot 10^{-100}$

(b) Spotting performance in terms of F_1 for different penalty parameterizations.

Figure 7.13: Evaluation and optimization of the syllable spotting mechanism.

6.1.4) given a partially learned syllable representation. Because of the lack of a sufficiently large poly-syllabic speech corpus, the evaluation corpus was created using the same approach as in the last section. Multi-syllable utterances were generated by the SYTU instrumented with distinct sets of test S_T and background S_B syllables.

All possible conditions for segment extraction can be condensed into three classes:

1. An utterance *starts* with a segment containing one or more not yet learned syllables.

2. An utterance *ends* with a segment containing one or more not yet learned syllables.

3. Two already acquired syllables *enclose* a segment, which contains one or more not yet learned syllables.

Thus, the structure of the generated utterances was statistically constrained to give instances of each class. Hereby, the number of syllables in background and test segments was sampled randomly from $\{1, 2\}$. This resulted in utterances with lengths between 2 and 6 syllables, which matches the utterance length of spontaneous speech.

Segmentation results as provided by the syllable spotter, became subsequently processed by the subtractive decomposition algorithm described in section 6.4.2. The used optimization function was chosen to be the number of correctly detected background segments.

We assessed the quality of the segment extraction for all three problem classes separately to differentiate problems with respect to the position of the embedded background segment. We tested training segment candidate generation by using supervised trained syllables models and the unsupervised bootstrapped background phone model pool $\mathcal{M}_P^W$. This experiment did not investigate the decomposition of extracted background segments into syllabic subsegments as this is assessed separately in section 7.3.6. Figure 7.14 visualizes the results for all three conditions, each evaluated with 500 test utterances. For each class, precision and recall for correctly detected background segments are reported in the left sub-figure. Hereby, a phone-sequence segment was considered to be a hit, when it matched the underlying background syllable segment with more than 75%. As shown in the figure, detected segments were in most cases actual hits. However, not all background segments were found under all three conditions. The system tends to perform well on initial and

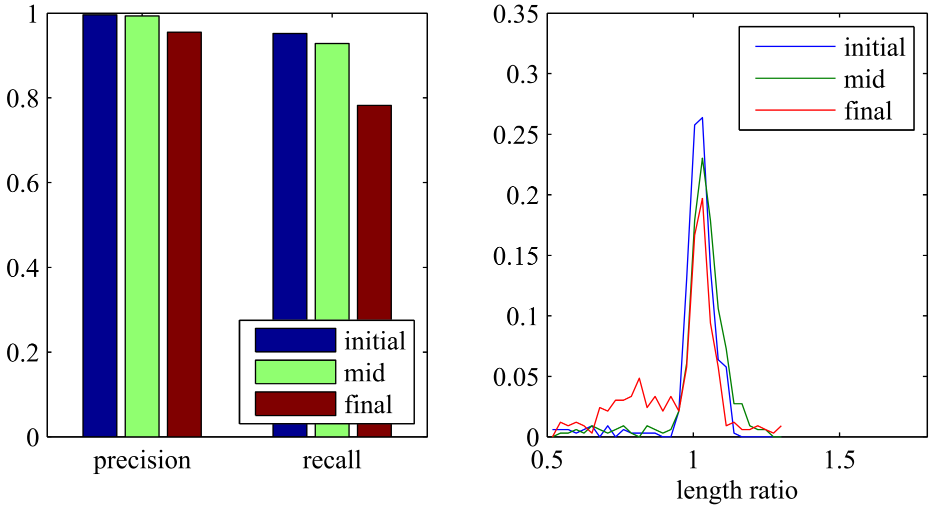


Figure 7.14: Performance of syllable training segment candidate extraction by applying the principle of subtraction.

wrapped background segments, but detects just around 80% of embedded background model syllables in the final position.

To assess the accuracy of the segmentation, the left sub-figure in fig. 7.14 depicts the average ratio of background syllable segment length and actually detected background segment length. In all three conditions the detected segment boundaries match with high precision to the ground truth data. By combining both findings, we conclude that the implementation of the proposed model is not able to detect all background syllables, but actually found segments match with high precision the actual structure of the input signal. With respect to the task of speech acquisition this is encouraging, as syllable model quality crucially depends on the segmentation accuracy.

7.4.5 Syllable Grammar learning

Syllable detection is constrained by an incrementally bootstrapped probabilistic syllable grammar L_S as discussed in section 6.4.5. To investigate the learning process properties, we first assessed the performance using semi-synthetic speech created by SyTu. Instead of the flat generative syllable grammar, a simple artificial syllable language grammar comprising 30 syllables depicted in 7.15 was used to constrain syllable sequence generation. This generative model SyTu$_{\text{ART}}$ had been designed to contain also some poly-syllabic constructs (e.g. `toy bird`, `take this` or `give me this`), which can be considered as poly-syllabic words of our artificial language (cf. 7.5.1).

The system was instrumented with a supervised trained syllable model comprising all 30 elements of the grammar. 2000 utterances became processed for evaluation. As discussed in section 6.4.5, detection hypotheses were employed to incrementally bootstrap L_S. As detected syllable sequences will include all types of errors discussed in section 7.1, these will propagate also into L_S. To assess the quality of the emerging syllable grammar we computed two statistics. First, we calculated the average length-normalized utterance L_S probabilities on a distinct set of test utterances generated with SyTu constrained either with the rule grammar 7.15 or with a flat-distributed grammar (both sets were not part of the L_S-bootstrapping sample). Second, we used the generative property of L_S to generate a random sample of 1000 utterances. This allowed to

```
#JSGF V1.0;

public <all> = /10/ <greeting> | /30/ <command> | /30/ <question> | /30/ <confirm>;

<greeting> = hello ( you | asimo );
<command> =   <action> [ <property> ] [<color> ] <object>;
<question> = [ where ] [ <property> ] [ <color> ] <object>;
<confirm> = ( yes | no ) [ <object> ];

<action> =  put | (give me this) | (show me) | bring | learn | point | (take this);
<property> =  nice | correct | wrong;
<object> = cup | car | mouse | (toy bird) |cloud;
<color> = green | white | (red small) | blue;
```

Figure 7.15: Rule-grammar used for syllable grammar learning evaluation. The specification of the used grammar format is documented in [Hun00]

calculate the ratio ξ_{art} of well-formed utterances with respect to grammar 7.15.

The results are summarized in figure 7.16. Starting with a flat transition model, the recognized syllable sequences allow to bootstrap L_S with high accuracy. This is especially supported by the progression of the grammar-reconstruction score that matches to the underlying grammar model by over 70%. As this performance is achieved without the application of further post-processing techniques as confidence filtering on the syllable segmentation hypotheses, further performance improvements seam to be feasible.

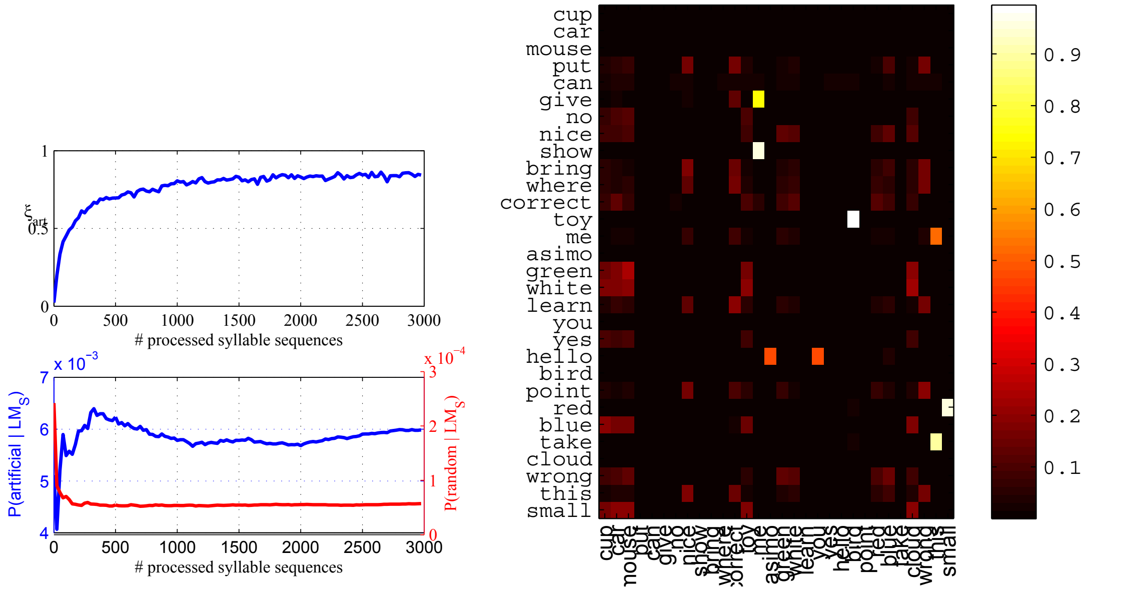
7.5 Words

Words define the highest level of speech perception in our proposed model. They need to be evaluated differently, as their representation differs considerably from that of phones or syllables. Because of the discrete nature of words in our model, most functionality of the word layer could be investigated using symbolic speech. However, to link up all layers we employ acoustic speech as input to our system, to validate whether the proposed model is actually able to reveal a word representation under realistic conditions. The complete hierarchy needs to be confirmed to segment acoustic entities of word length.

7.5.1 Lexical Learning

We evaluated the word acquisition properties using SyTu as input. The generative model of SyTu was chosen again to be the artificial language 7.15 (referred to as SyTu_{ART}). The experiment was conducted by incorporating the probabilistic syllable grammar L_S of section 7.4.5 learned from a semi-synthetic acoustic speech signal. The focus of the experiment was to investigate whether the lexical acquisition approach 6.11 is able to reveal the word structure of of the artificial language 7.15. By using the same experimental setup as above, we now process syllable sequences not only to refine L_S but also as input to the word layer described in section 6.5. The base threshold for poly-syllabic word learning Θ was set to $\Theta_0 = 2.5$ and the delay term to $d_0 = 1000$.

As the statistics of the underlying language are known, we could assess the performance in terms of the number of correctly identified mono-syllabic and poly-syllabic words. The test-language was designed to contain mostly monosyllabic and only few bi/tri-syllabic constructs. This was to match



(a) Grammar-matching ratio and average length-normalized utterance probability progression during the bootstrapping of L_S.

(b) Transition matrix of L_S. The bi-syllabic constructs are clearly separated from the other syllable transitions.

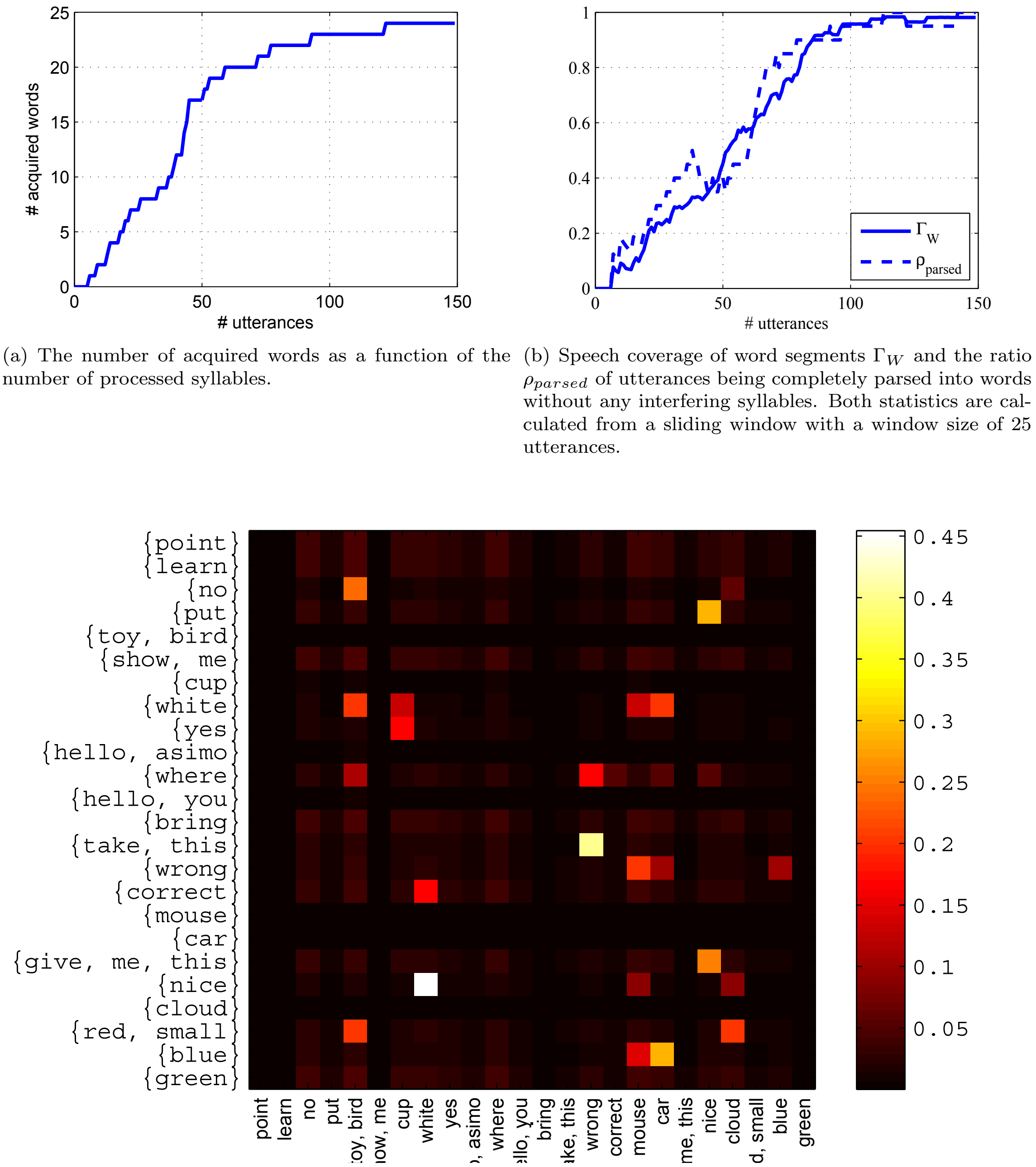
Figure 7.16: Incremental learning of the syllabic transition structure. The input to the system were sequences of actually recognized syllables from the semi-synthetic speech signal generated with SyTu which was configured with the grammar model 7.15

the statistics of most real languages which contain a magnitude more shorter than longer words (cf. [Gre98]).

Figure 7.5.1 depicts the acquisition process in terms of the number of correct words acquired, the number of incorrect words acquired, and the number of utterances successfully parsed. The system acquired the following words in the respective order : `{no}`, `{hello, asimo}`, `{yes}`, `{mouse}`, `{toy, bird}`, `{wrong}`, `{cloud}`, `{hello, you}`, `{nice}`, `{where}`, `{give, me, this}`, `{point}`, `{bring}`, `{cup}`, `{car}`, `{learn}`, `{white}`, `{blue}`, `{green}`, `{correct}`, `{red, small}`, `{put}`, `{take, this}`, `{show, me}`. The learned word representation almost perfectly matches the structure of the artificial language 7.15.

The syllable grammar L_S successfully enabled the system to reveal bi- and tri-syllabic syllable patterns as words. As the number of monosyllabic utterances generated by SyTu$_{\text{ART}}$ is very low compared to longer utterances, the lexical acquisition relies to large parts on the self-referential processing scheme (step 3 in alg. 6.11). Especially as action words do not appear as it, the cascaded learning sequence is necessary to extract them from the acoustic input speech. This is also reflected in the learning order as shown in the listing: action words tend to be learned after property and object words have been acquired.

Figure 7.17(a) depicts the speed of word learning. The frequency of new word model creations decreases quickly after an initial phase of very active learning. This is because residual learning allows to reveal the word structure quickly as soon as the first words have been acquired. From figure 7.17(b) it becomes clear, that the word representation allows to parse the complete syllable input sequence as decoded by the syllable spotter from the acoustic input signal. Finally, the incrementally learned word bi-gram L_W is depicted in 7.17(c). It clearly reflects the syntax model used by SyTu$_{\text{ART}}$.

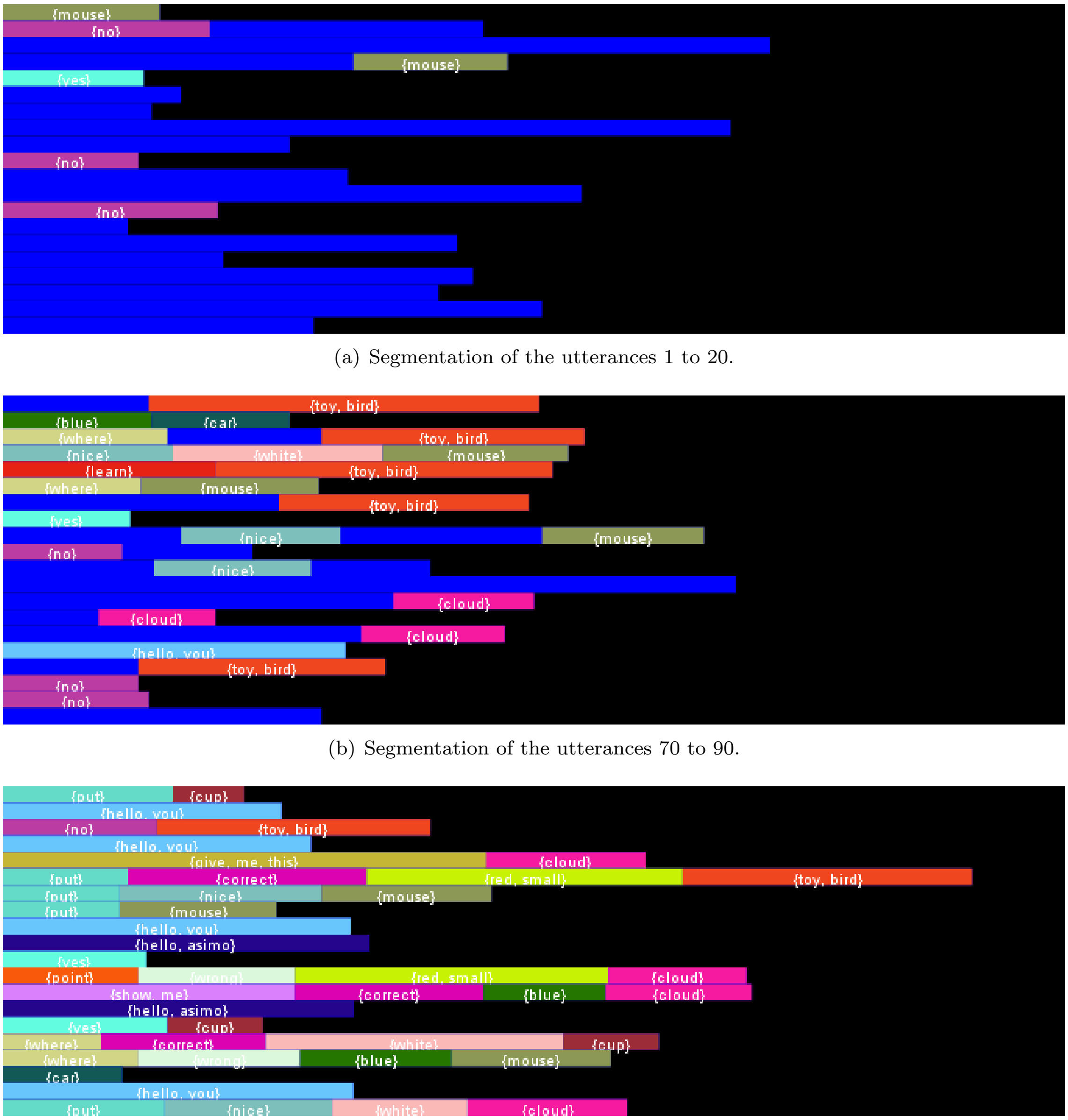


(a) The number of acquired words as a function of the number of processed syllables.

(b) Speech coverage of word segments Γ_W and the ratio ρ_{parsed} of utterances being completely parsed into words without any interfering syllables. Both statistics are calculated from a sliding window with a window size of 25 utterances.

(c) Transition matrix of the incrementally trained word bi-gram L_W at the end of the experiment.

Figure 7.17: Lexical learning using acoustic speech as input generated by SyTu$_{\text{ART}}$



(a) Segmentation of the utterances 1 to 20.

(b) Segmentation of the utterances 70 to 90.

(c) Segmentation of the utterances 130 to 150.

Figure 7.18: Segmentation snapshots at different time instances. Filler syllables are displayed as dark blue. As the number of acquired words increases, the word parsing improves considerably. Even if not present in the depicted examples, artifact syllable segments occasionally distort the input of word layer. They are part of the input, as phones were not included as filler-model for the experiment. This was because all syllables produced by SyTuArt were available as supervised trained models to the syllable spotter. Such artifacts can be partially avoided by using a higher syllable insertion penalty, as investigated in section 7.4.3.

CHAPTER 8

Embodied speech acquisition

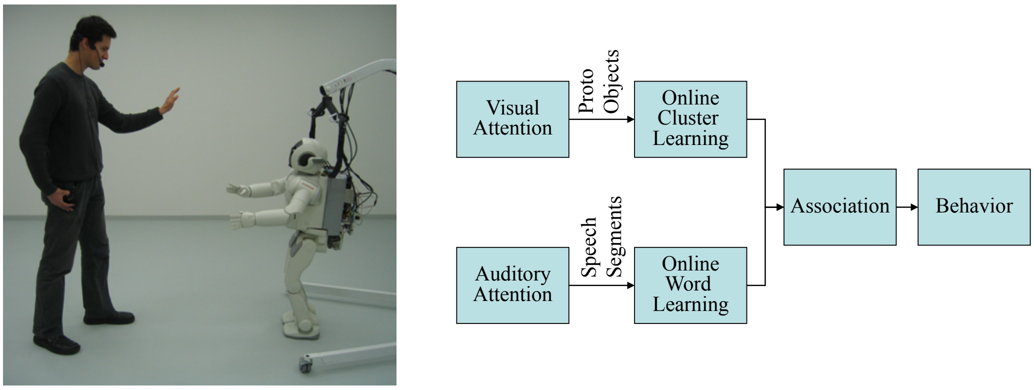
According to Hickok [Hic09] speech perception is best conceptualized as a process that allows a listener to access a lexical concept - such as the meaning of a word - from a speech signal. So far we have investigated speech structure acquisition decoupled from any embodiment. Even if our model aims to be a computational model of the former process, it is nevertheless mandatory to show that it also embeds naturally into an embodied context. This includes two major aspects. First, it is necessary to show that acquired word entities can be linked to semantic categories. Second, as speech production and perception are considered to be deeply interrelated to each other, we need to investigate how such an integration could look like. Thereby we keep our focus on developmentally plausible processing and interaction schemes.

8.1 Grounded word acquisition

The focus of this work is on developmentally inspired models for acoustic language acquisition. However, as these require some kind of embodiment in order to ground acquired words entities, we have embedded the speech structure acquisition model proposed in chapter 6 into an *a*utonomous multi-modal *l*earning and *i*nteraction *s*ystem ALIS[Goe07] [Mik08] [Bol08] [Goe09] running on Honda's ASIMO robot as depicted in figure 8.1(a).

Inspired by developmental principles, we aim to build systems that develop cognitive abilities in a childlike manner. But as the development of infants spans years it is hardly feasible to present computational models within a robotic demonstration scenario. As we believe learning to be grounded in interaction with a tutor, it is not applicable to include offline-learning into a demonstration to leap periods in which the robot does not show any significant progress in cognitive abilities. A possible approach which seems to be much more feasible to us, would be a snapshot-function on system level. This would allow to prepare certain snapshots where the developmental state of certain cognitive function could be demonstrated. However, such an approach would restrict demonstrations to states. The underlying bootstrapping processes which are the main focus of our research would be completely neglected.

Thus, we followed a simplified scheme in ALIS which was to equip ASIMO with the ability to learn new auditory labels online in interaction. Even if we embedded the complete model of section 6 into the ALIS system, it was evaluated in a simplified configuration. The strongest simplification was the inbuilt assumption that training utterances contain only one word, each modeled as a distinct syllable model and a respective monosyllabic word model.



(a) Interaction between tutor and robot. Here the first iteration of the system is shown, where Asimo was not yet equipped with an auditory attention system. The tutor needs to use a (mutable) head-set microphone to talk to it.

(b) Interaction-driven semantic learning with an embodied system. Only the speech-relevant elements are depicted. For details see [Bol08]

Figure 8.1: Autonomous semantic learning in interaction with a humanoid robot.

A key principle behind the multi-modal learning processes is commonly referred to as *cross-situational learning* [Pin89]. It is grounded in the idea that infants cannot observe one-to-one correspondences between words and referred entities but rather analyze large amounts of utterances to defer the meaning of single words. For instance, the word `yellow` cannot be grounded from a single scene containing a yellow duck. Only after a multitude of sceneries containing yellow objects accompanied by a scene description containing the word `yellow`, infants are able to reveal the words' meaning. This learning process can be implemented using EM estimation techniques, where the associations between semantic entities and languages referees are represented by a probabilistic model [Duy02] [Bal03].

In our system, neither acoustic word labels nor object property clusters are innate. The system starts with an empty multi-modal representation. Hence, according to the design of the proposed model the robot is not able to learn any auditory labels, but simply accumulates statistics about the language. These are subsequently condensed into a set of phone models as described in section 6.3. Independently of a concrete appearance the system is innately able to detect object motion, size, planarity, and object location relative to the robot's upper body.

As words are treated as syllables in ALIS, word learning works as follows: Given an object, the user restricts the system's attention to an object property which should be labeled (e.g. an object's height). This can be regarded as a bias towards the semantic class the tutor wants to teach a label for. By doing so we completely avoided the problem of cross-situational learning, and made the system demonstrable also with only a few training examples.

New labels are then taught by providing a few (2-5) isolated samples of each word. While addressing these training samples to the robot, the tutor presents the realization of the object property she wants to teach to Asimo. The temporal grouping of these speech samples is given to the system as an additional cue to ease the syllable clustering process. The temporal learning window becomes disposed when the tutor removes the object from the robot's visual field. Subsequently, speech and non-speech representation are updated. This is either achieved by updating an existing cluster or by creating a new cluster in the respective sub-representation. For instance when teaching `top` while indicating top positions with an arbitrary object, the robot will learn two clusters (word, position model) linked via an associative model. The basic elements of the system are depicted in figure 8.1(b).

To evaluate learned semantics an arbitrary object is presented to the robot. Then an auditory label is addressed to it. The robot nods, if the perceived object properties match the tutor's description. Otherwise it shakes his head, but keeps the expectation towards the uttered auditory label. Subsequently the tutor can modulate the object's properties (for instance by moving it to another location) to fulfill the expectation. Non-matched expectations are deceased after a fixed time-interval.

Based on the novelty detection mechanism presented in 6.4 the system is able to distinguish automatically between already known synonyms and new synonyms. The same mechanism allows to retrain already learned synonyms to improve the recognition performance. The system is by design language-independent and was also successfully used to acquire mixed-language representations comprising words from up to 4 languages. In our experiments up to 30 words could be learned online through contingent verbal and gestural interactions. No offline computation was necessary and almost no words were confused (cf. [Mik08] and [Sch05] for details).

In the first iteration of the system [Goe07], we relied on a headset for speech input as depicted in figure 8.1(a). This was necessary to avoid background speech to disturb an ongoing interaction. Even if such a technical solution is widespread in human-robot interaction [Iwa03] [Bal03] [Roy00] it lacks of naturalness and prevents additional tutors to join an ongoing interaction. Thus, we extended our system with an auditory extension system as proposed in [Hec09]. The basic idea was to restrict the auditory attention of ASIMO to situations when it is actually in interaction with a tutor. This enabled a free interaction with our robot without the need to use a headset.

To make this approach functional it was crucial to compensate the ego-motion noise and the background noise in the lab. However, even with adaptive noise cancellation methods (cf. [Hec09] for details), the input signal was still considerably more noisy compared to head-set microphone recordings. As MFCC features used throughout this thesis are known to be less noise-robust we changed the feature space to biologically-inspired spectro-temporal HIST features as proposed in [Dom09]. By doing we were able to interact with ASIMO also in noisy environments. This proved that our model is independent of a specific type of speech representation.

8.2 Linking speech perception and production

Social interaction between human and robot requires the robot to understand and to produce language. But these faculties are by no means trivial and need to develop in interaction with a caregiver. Thereby, speech understanding vastly outperforms the early infant's ability to speak. Usually infants start to utter their first one-word utterances at the age of 10 to 11 months. At this age they may understand already between 10 and 100 words [dBB99]. Typically speech production is preceded by re-duplicative *babbling* around 6 to 10 months after birth. It is composed of repetitive syllables patterns like `bababa` or `mimimi`.

By the age of around 18 months infants start to produce first 2-word phrases. Speech at this age still lacks of grammatical inflections, suffixes and function words. However, infants at this developmental stage rarely make word order errors (cf. [Kit03]). This indicates that they have

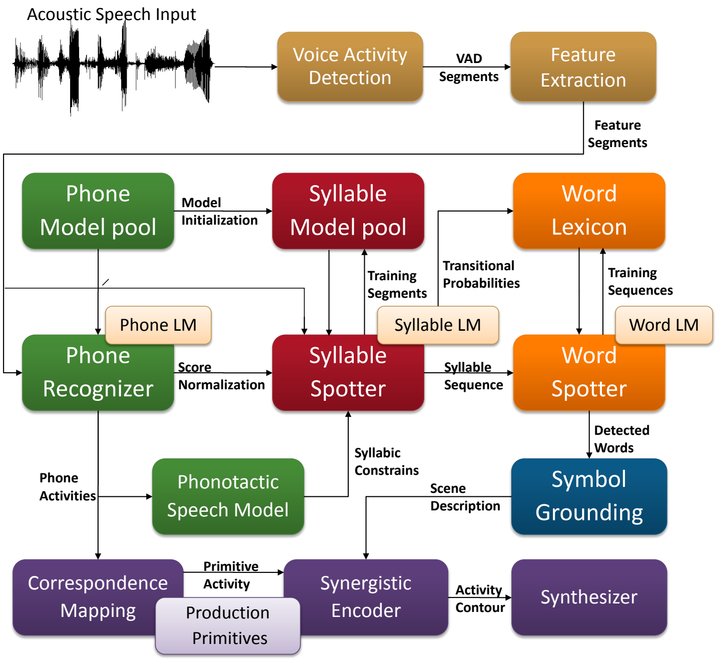


Figure 8.2: The extended architecture for coupled speech acquisition and production. The coupling between production and perception takes place mainly at the level of the phones and motor primitives that are linked through a correspondence model. This allows detected phone sequences to be mapped to vocal primitive activities, which are converted by the synergistic encoder into a continuous sequence of vocal commands, which becomes then realized by the speech synthesizer.

already acquired a basic but functional syntax model of their language.

Although it is out of scope of this thesis to fully extend the proposed model for speech structure acquisition to account also for the marvelous processes of speech production, we were at least interested to see whether and how our model could embed into such an integrated system. Because our approach defines a bootstrapping process in the acoustic space it seems to be natural to link it to speech production. Thus, we extended the model proposed in this thesis with a scheme to equip our robot with the ability to imitate its tutor. This has been achieved by integrating our model with a system for speech imitation learning started by Vaz in [Vaz09c].

The extended model [Vaz08] [Vaz09c] [Vaz09b] implements a tight coupling of perception and production, namely a correspondence model between phones and motor primitives innate to our robot. This coupling is built through an exploratory process, in which the system learns the consequences of its vocal actions, in terms of the tutor's voice. Using statistical inference, our system allows to convert a tutor utterance into a probabilistic sequence over the system's vocal repertoire, that becomes subsequently transformed into a synergistic motor coding, used to imitate the tutor utterance. To evaluate this integrated speech acquisition and production model, we present an interaction experiment between a human tutor and our robot.

8.2.1 Speech production architecture

Speech imitation layer

The ability to find own motor configurations that produce phonetic equivalents of words uttered by a tutor is crucial to speech imitation. This is by no means a trivial task, due to significant differences between the voices of the caregiver and the infant. Different lengths and proportions of their vocal tracts cause these differences, which include different pitch and formant frequencies.

The extended architecture depicted 8.2 allows to learns a correspondence mapping between vowel sounds generated by a system innately equipped with a child's voice and the equivalent vowels from its tutor. Thereby no assumptions on the language of interaction or the phonetic properties of the tutor or system's voice are being made, and learning relies solely on an imitative response of the tutor. The imitation subsystem consists of three main modules:

- A probabilistic mapping between the phone models and the system's vocal repertoire
- A synergistic encoder that converts a probabilistic distribution over the set of motor primitives into a motor activation
- A synthesis module that synthesizes motor activations into an acoustic speech signal

The correspondence mapping $\mathcal{C}$ represents for each phone model λ_i^p a probability distribution over the space of motor primitives, which are denoted to as m_j respectively.

$$\mathcal{C}_{ij} = P(m_j|\lambda_i^p) \tag{8.1}$$

In section 8.2.2 we describe how this mapping is learned.

Synergistic encoder

Given an utterance to be imitated, the synergistic encoder computes a motor output from a sequence of recognized phone model hypotheses provided by the phone recognizer (cf. figure 8.2). Each segment hypothesis has an associated time span $[t_0, t_1]$, and becomes converted into a distribution over all motor primitives as encoded in the correspondence mapping $\mathcal{C}$. The synergistic encoder interprets these probabilities as activations for each motor primitive.

The set of motor primitives and associated weights are transformed into a continuous output function by means of a morphing function. For every phone segment hypothesis, we assign each motor primitive an activation contour whose strength depends on its weight. This activation contour is Gaussian with a variance dependent on the duration of the segment. The overall activation of each motor primitive is calculated by the summing over the activation contours of all segment hypotheses.

$$\mathcal{W}(m_j) = \sum_{\lambda_i^p} \mathcal{C}_{ij}\, G((t_1 + t_0)/2, k(t_1 - t_0)) \tag{8.2}$$

Synthesis module

A vocoder-like scheme is used to synthesis the actual imitation speech signal. This scheme is suited to synthesize complete sets of phonemes for many types of voices, especially high-pitched ones like those from children. In spite of recent developments, articulatory models do not offer that possibility. For this synthesizer, motor primitives have the form of spectral vectors, annotated

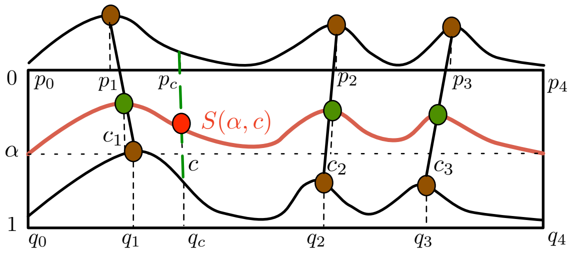


Figure 8.3: The morphing algorithm computes the spectral value $S(\alpha, c)$ for each channel c at an intermediary position α given initial and final spectral vectors p and q, and a correspondence matrix associating (p_i, q_i).

with their formant frequencies for use of a morphing algorithm. The set of motor primitives and their activations is transformed into a unique spectral output with a speech morphing algorithm, and becomes subsequently synthesized into a waveform.

Spectral morphing algorithm

The overall activation from equation 8.2 is translated into a single output by means of a morphing operation. Morphing two spectral vectors results in a third spectral vector representing an intermediate state, where the value of each channel is given by

$$\mathcal{M}(m_j, m_k, \alpha_j, c) \quad = \quad (1 - \alpha_j)\, m_j(p_c) + \alpha\, m_k(q_c) \tag{8.3}$$

$$\alpha_j \quad = \quad \frac{\mathcal{W}(m_j)}{\mathcal{W}(m_j) + \mathcal{W}(m_k)} \tag{8.4}$$

Here, $m_j(c)$ refers to channel c of m_j, and p_c and q_c are calculated by maintaining the proportion of the distance from channel c to the immediately inferior q_c, respectively part of the initial and final spectral vectors.

Synthesis algorithm

Spectral vectors are synthesized into speech using an algorithm based on the channel VOCODER, see [Vaz09c] for more details. This algorithm was developed in order to allow the synthesis of children's voices. It makes use of a gamma-tone filter bank at its core, which allows for an optimal trade-off between spectral and temporal resolution. As a consequence, high and low pitched voices can be synthesized with similar quality and without the need of any special speaker-dependent model.

8.2.2 Interaction

The interaction scheme necessary for imitation learning naturally extends the multi-modal semantic learning framework described in section 8.1. From our experience with the system, it became clear that for a more natural interaction simple gestures like head nodding are not sufficient. Speech production abilities are crucially for such a system as these equip the robot with the ability to reflect what it has learned in interaction. For example, while presenting a red apple on the right side, our system should be able to provide an acoustic scene description like "right red apple".

For this, the system needs to be able to project the acoustic targets of the learned labels into its own articulatory space. In our system, this correspondence model $\mathcal{C}$ is encoded on the phonemic level, and is learned through interaction with the tutor. We integrated this learning process in the

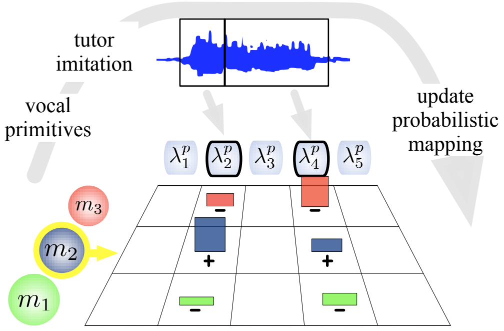


Figure 8.4: Example for correspondence model learning. A randomly picked vocal primitive m_2 is synthesized with constant timbre. The tutor imitates the vowel sound and the response is used to update the experience mapping in proportion to the amount of activation of each phone.

overall interaction scheme, by making the imitation sub-system to initiate interaction after a given period of inactivity: The system produces one of its basic vocalic sounds, and uses the tutor's imitative response to refine the probabilistic correspondence mapping.

Tutor imitates system

In this training phase the system learns a correspondence between its articulated motor primitives and the imitative responses from its tutor. Vowel primitive utterances are produced with constant timbre by synthesizing spectral vectors from its repertoire. The tutor then imitates the system, which determines the best matching phone sequence

$$[\lambda_1^p, ..., \lambda_n^p] = \arg \max_{[\lambda^p] \in \mathcal{P}} P([\lambda^p]|X_{tutor}) \tag{8.5}$$

The experience mapping M representing the probability of perceiving phone model λ_i^p given a vocal primitive m_j

$$M_{ij} = P(\lambda_i^p | m_j, Dj) \tag{8.6}$$

is then updated in proportion to the segment length of each detected phone model λ_i^p. Hereby, $\mathcal{P}$ denotes the set of all possible phone sequences, $[\lambda^p]$ a sequence of phone models, and X_{tutor} indicates an acoustic speech sample from the tutor. This procedure is schematized in figure 8.4.

System imitates the tutor

In order to imitate, the system maps phone model likelihoods to activations of vocal actions, using the probabilistic correspondence mapping described in equation 8.2.

The correspondence mapping is inferred from the experience mapping, see equation 8.6.

$$C_{ij} = P(m_j|\lambda_i^p) = \frac{P(\lambda_i^p|m_j, Dj)\, P(\lambda_i^p)}{P(m_j)} \tag{8.7}$$

Because we assume a flat prior over all motor primitives, and the values of the mapping are only computed considering a single phone model at a time, the correspondence mapping can be represented as

$$C_{ij} = M_{ij} \tag{8.8}$$

The likelihood of each motor primitive is passed onto the synergistic encoder, which computes a time sequence of motor primitive activations as described in section 8.2.1. This sequence is

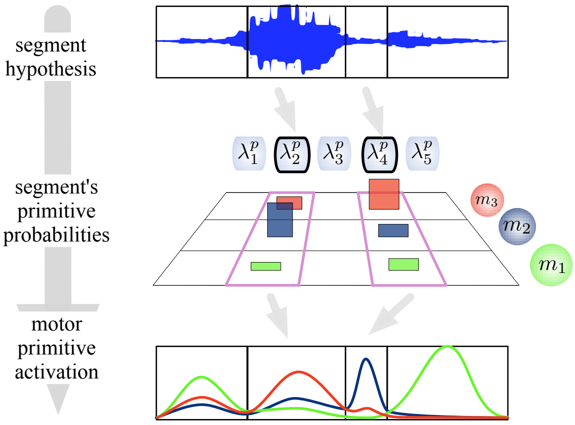

segment hypothesis
segment's primitive probabilities
motor primitive activation
λ_1^p λ_2^p λ_3^p λ_4^p λ_5^p
m_3
m_2
m_1

Figure 8.5: A tutor utterance to be imitated is parsed into a sequence phone segments. Using the correspondence mapping, the probability of each motor primitive for the phone models most active in the different segments is computed and used by the synergistic encoder to generate a motor activation.

then recursively morphed into a single vocal output for each time instant, according to the motor primitives' relative strength of activation, and passed onto the synthesis algorithm that generates the imitation signal.

8.2.3 Experimental results

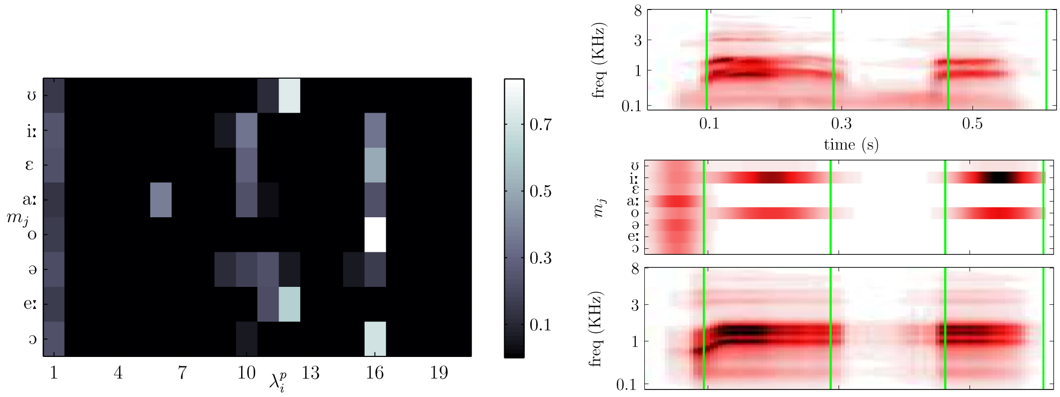
Given a learned phone representation, we evaluated the correspondence model bootstrapping as described in section 8.2.2. We grounded the system's voice in a set of 8 spectral vectors, selected from cluster centers computed with the K-Means algorithm over the spectrograms of utterances spoken by a 10 year old male child, from the TDIGITS corpus [Leo84]. Each spectral vector comprising 100 channels with center frequencies from 40Hz to 8KHz, represents one of the following vowels (IPA alphabet): ɔ, e, ə, o, a, ɛ, i, ʊ.

Each of the vocal primitives was synthesized and played 15 times to a male adult speaker, who imitated them. We synthesized each robot's utterance with a random duration (between 0.25 to 0.3 seconds), and different pitch contours. The resulting correspondence model is depicted in figure 8.6(a).

The following aspects can be observed from the data. Firstly the imitative response of the tutor only covers a subset of the set of phone models. This was expected, because the system is limited to the production of vowel sounds, and the phone models are trained using unconstrained speech containing both vowels and consonants.

Secondly, phone model with index 1 has a very strong response for all tutor responses; this is an artifact due to our voice activity detection that includes short noise parts in the beginning and at the end of the detected speech segments.

Thirdly, the models for the different vocal primitives vary considerably: primitives for vowels ɔ, e, ʊ have a very unimodal response, while others like e have a more disperse response. Several factors might be contributing to this, the most likely being either a non-uniform imitative response of the tutor to the vocal primitive or the in-existence of any phone model fully representing the imitative response. One reason supporting the first might be that, although the vocal primitives were selected with care to correspond to one vowel, synthesizing a sound with constant timbre presents limitations to its naturalness, not necessarily affecting all vowel sounds equally. One reason supporting the second is that the phone models are trained using different data, even if



(a) An instance of a correspondence model learned in interaction with a tutor. The phone OPDFs conditioned with the different motor primitives are shown in the rows. The vowel phones (represented using IPA notation) are marked.

(b) An example of an utterance imitation performed by our system. Depicted are (from top to bottom) enhanced input utterance spectrogram, the resulting motor primitive activations, and the output spectrum used for synthesis.

Figure 8.6: Examples for the correspondence model and an imitated utterance.

originating from the same speaker. We tried to compensate this effect by balancing the words in the training corpus according to the vowels contained, but issues with over- or under-representation are seldom avoided in unsupervised learning systems. Another possibility would be to (at least partially) overlap the phone model learning phase with the learning of the correspondence model, so that the phone models can be estimated using similar data. The disadvantage would be that the interaction phase would take longer.

An example of the different stages of processing can be seen in figure 8.6(b), where the word *mama* is imitated. As already explained, only the vowel segments are being imitated. Thus, the words the system produces can be distinguished if the vowel constituent's sequence is different.

CHAPTER 9

Summary and discussion

Finding structure in spoken language is a central problem in computational linguistics. It is not yet understood how infants master this challenge. Findings from developmental psychology indicate the ages at which infants are able to succeed in which cognitive tasks. This gives at least a rough idea about how speech perception evolves over time. Because of the diverseness of speech segmentation cues it is reasonable to assume that no single factor alone accounts for the development of the ability to segment speech into words.

To our best knowledge previous computational models offered only very limited explanation for this marvelous process observed in infants. We believe this to be mainly caused by a too strong focus on single speech granularity levels and symbolic speech as input. It is still an open question how infants select appropriate segmentation strategies depending on the speech context and their sensitivity to different segmentation cues. This defined our starting point and major motivation while working towards a computational model for speech acquisition.

Inspired by processing principles which are believed to play a role in early infant speech development, we have proposed a novel model for unsupervised acoustic speech structure acquisition. It links phones, syllables and words by implementing developmentally plausible processing schemes. Our model can be applied to arbitrary languages and generalizes also to languages with a syllable structure that differs from the linguistic definition in terms of onset-nucleus-coda as discussed in chapter 2.3. The only assumption here is, that certain phonotactic principles restrict the constitution of syllable-like units. The proposed model contributes conceptually, methodologically and experimentally to the field of computational speech acquisition. We conducted experiments to validate function and performance of the different sub-processes, and provided new insights into the integrated learning of speech structure. We could show that our current system is able to learn a stable set of syllable and word models independently of the complexity of the test language. Clearly, because of limited resources we were able to validate the model only to some extent. But given the complexity of the problem, the lack of standardized benchmarks for most of the different subsystems, and only limited access to appropriate evaluation corpora, we consider the achieved results to be promising.

Phonotactic learning for acoustic speech

We could show that the proposed system is able to bootstrap a phone representation in a solely data-driven manner. Word recognition results obtained by using distributed phone-sequence models were not comparable to results using common speech recognition methods (cf. sec. 7.3.3). This confirmed our initial assumption that co-articulation between adjacent phones does not allow a strictly hierarchical organization of speech perception.

Inspired by ideas implemented for symbolic syllable structure acquisition models, we proposed a novel method to learn phonotactic constraints from phone sequences as detected in acoustic speech. This we modeled using a set of coupled incrementally bootstrapped phone transition models that were designed to capture the syllabic structure from the utterance boundaries. The method was shown to give meaningful clues for syllable boundary prediction.

Training segment generation

Based on this phonotactic model, we proposed a phonotactic parser to decompose speech segments into syllabic units. We confirmed experimentally that this parser gives reasonable decompositions (cf. sec. 7.3.6). A perfect decomposition by solely employing this parser has not been achieved. However, this is coherent with the finding that model-based speech segmentation methods tend to outperform less informed classifiers. This clearly applies for the problem of syllable segmentation, which we further investigated with incrementally learned syllable models. Thus, phonotactic parsing can be thought of as a seeding mechanism that provides cues to determine syllabicity in a local context.

This interpretation allowed for the next major contribution of this work: a combined scheme to derive training segments for syllable learning from increasingly more complex utterances. This we modeled by means of a residual-learning scheme given the results of a syllable detector combined with the phonotactic parser. We validated that such an approach is able to reveal training segments with high segment boundary precision, especially under the assumption of an incomplete syllable representation (cf. sec. 7.4.4).

Incremental syllable clustering

Whereas our initial motivation was to develop a model for acoustic word learning and segmentation, it became clear that syllables define the unifying concept of most word segmentation principles. But because syllables cannot be considered to be innate, any computational model for lexical acoustic acquisition needs to cope with the problem of syllable learning in beforehand.

We proposed a novel incremental syllable clustering method, which we consider to be the most important contribution of this work. It implements a divisive clustering scheme to reveal the syllable structure of an arbitrary input language. The result of this process is a syllable representation, that comprises distinct HMMs for each syllable. The system acquires syllable models incrementally as they appear in time. To make it functional we proposed a regulatory framework of local and global control that modulates the parameters of the bootstrapping process. Additionally, we proposed a novel initialization method for syllable unit HMMs, that we have shown to outperform existing approaches (cf. 7.4.1).

We were able to show that this syllable clustering reveals a highly discriminative set of syllable models independent of the number of syllables contained in the input language (cf. sec. 7.4.2). Furthermore, we validated in our experiments that this process is capable to detect the number of syllables in the input language with high confidence (cf. sec. 7.4.2). Finally, we showed that the process successfully trades stability against plasticity, as it is able to cope with changing syllable input (cf. sec. 7.4.2).

Lexical learning

To reveal words from detected syllable sequences, we proposed a novel lexical learning algorithm. It was designed to integrate assumptions about child-directed speech, co-occurrence learning using the incrementally learned syllable transition model, and a self-referential scheme to decompose utterances into known and unknown segments.

We validated this lexical learning algorithm on an artificial language designed to include polysyllabic words as well as words that could only be acquired by means of residual learning. We showed that the method was able to acquire the correct word structure of our test language. We also could show that the incrementally learned word transition model allows the reconstruction of the syntax of the input language with increasing confidence (cf. sec. 7.5.1).

Embodied semantic learning

It was out of scope of this thesis to fully bridge the gap between unimodal speech structure acquisition and the way infants learn language through a multitude of complex multi-modal interaction patterns. However, in contrast to most other works on speech acquisition we successfully embodied our system into a large-scale multi-modal semantic learning architecture running on Honda's humanoid robot Asimo. This enabled us to prove that the proposed model allows incremental online learning of semantics in interaction with a tutor by linking meaning to word symbols provided by our system. The integration showed that our proposed model for speech acquisition embeds naturally into a semantic learning process.

It is clear that speech is not bound to the acoustic speech signal but is rather a potpourri of all means of human abilities to express themselves. This includes for instance mimics, posture, blinking and gestures. Hence, we consider the embodiment of the proposed model as a first step to account for those effects. For instance we observed much more natural speech intonation patterns when addressing the robot compared to the rather artificial recording setup in front of desktop computer.

Integration with imitation learning framework

We presented and tested an integrated approach for infant-inspired speech acquisition and production by coupling the proposed embodied speech acquisition model with an imitation system. We showed how to link the phone representation to the motor primitive space of an imitation system. This we achieved by means of a probabilistic correspondence mapping learned in interaction with a tutor. By assuming a cooperative tutor that imitates monophonic utterances of the system, we proposed how to bootstrap such a correspondence model of the tutor's imitative response to each of the system's motor primitives.

As the result of this tutoring process, the system was able to imitate voiced words with its own voice. Specifically we proposed how such a model equips a humanoid robot with the ability to describe its environment in terms of labels for various object properties that have been associated to arbitrary words in interaction with a tutor.

Our approach extends previous attempts [Vaz09c] for sensory-motor coupling as it involves more and more plausible training data to estimate the perceptual part of the system: phones are learned not only from isolated phone-instances collected while learning the correspondence model, but from the complete interaction with the tutor. Although the vocal repertoire of the system contains only vowels, which obviously impairs the imitation of words containing consonantal sounds, we consider this to be an important step towards embodied developmentally plausible interaction-driven learning of speech production abilities.

9.1 Discussion and outlook

When developing a computational model it is necessary to maintain a balance between system complexity and limited resources required for implementation and evaluation. As a consequence some aspects could have been modeled in an alternate and possibly improved manner. However, as discussed above human speech acquisition is still far from being completely understood, and such additional mechanisms should be considered more as possible options than as better or even ultimate answers.

Inter-layer top-down feedback in addition to bottom-up processing

According to our understanding of infant speech development, a bottom-up organization of speech perception and acquisition seems natural. However, bottom-up bootstrapping of speech representation is possibly not a sufficiently rich model to provide a fully functional model of speech development. For instance Wally [Wal93] suggested that infants might represent words in a rather holistic manner. He further noted that infants may restructure their speech representation into a segmental model not before their vocabulary has reached a sufficient size. However, experimental evidence that supports this hypothesis is rare. In contrast 24-month-old infants were found in [Fer98] to initiate a saccade towards the respective item before the acoustic offset of a corresponding word. Clearly, this supports rather a bottom-up than a holistic organization of speech representation. Hence, as there are only few findings that actually support a top-down organization of speech perception, the bottom-up paradigm has been assumed and implemented in most computational models. But as no existing model - including the one presented here - was yet shown to provide a full explanation of speech development, top-down processing may be a missing key element.

In this thesis we presented a purely bottom up acquisition and processing framework. We included intra-layer top-down feedback by constraining speech unit decoding with incrementally learned speech unit transition penalties. Not evaluated in greater conceptual and experimental detail was an *inter-layer top-down propagation of contextual knowledge*, even if it seems reasonable to assume such a feedback to be beneficial to the performance of an emerging speech representation.

The proposed model captures phonotactic constraints from utterance boundaries only. This is due to the fact that without any knowledge about the syllable structure, syllable boundaries cannot be inferred from within an utterance. Hence, a natural extension (that is possible *only* because of the layered design of the proposed model) would be to employ syllable segments detected with high confidence as further training samples for the learning of phonotactics. Alternatively (or in addition), boundaries sequences of phone background sub-segments could help to further refine the phonotactic model. This would especially allow for capturing of syllabic constraints even of syllables that do not appear in utterance boundary positions because of the language syntax model. Such an approach might help to capture a more robust and more rich set of phonotactic constraints, as constraints learned solely from utterance boundaries may not generalize to all types of intra-utterance syllables (cf. [Chr98]).

However, it is not clear to us, whether such a kind of top-down flow of information would actually lead to an improved phonotactic parsing performance because the incrementally bootstrapped syllable representation is not necessarily as robust as utterance boundaries detected from speech activity contours. Thus, such an approach might rather flatten the phonotactic distribution than increasing its peakiness.

Another possibility to incorporate top-down feedback would be to employ the acquired *word model structure as a bias for syllable detection.* This would be straightforward to realize by incorporating additional transition penalties while compiling the syllable search graph. However, as for phonotactics it is not clear whether and how an incomplete and partially unstable word representation is suitable to bias more basic bootstrapping processes. But as syllable detection defines a central core mechanism of the proposed model, possible performance improvements by using such a top-down bias should be subject of further research.

Improved learning of phones and phonotactics

As phones are the basic units of speech perception in our model, it is clear that even slight improvements in phone processing may accumulate to significant improvements in syllable and hence word acquisition performance.

A striking way to improve phone clustering would be to tighten the link to speech imitation learning. While learning the correspondence model, the system assumes a cooperative tutor that imitates simple phone tuples (or single phones as in our implementation). Thus, the system is aware of the phone structure of the tutor's imitative response. Accordingly it could incorporate this knowledge into the phone learning process. This would allow to estimate very reliable clusters for all those phones that are part of the correspondence learning process. This could give a set of *seed phones*, and similarly to syllables, remaining phones could be subsequently derived by means of a self-referential learning process.

Direct speech segmentation approaches as discussed in section 5.2 are unlikely to result in phone segments that reflect the linguistic structure of the tutoring language. However, from our understanding of clustering processes we would assume such approaches to beneficially complement the proposed phone learning scheme. Especially we consider the incorporation of local change functions to be most promising as these provide an arbitrary fine-granular segmentation of the speech signal. Such a segmentation could serve as a non-flat prior for the clustering process, and would allow to shift the initial phone clustering process from a frame level up to a segment level: instead of speech features frames it would be become possible to cluster the segments directly. As this takes the adjacency structure of feature frames into account, we assume such a combined approach to improve phone model quality significantly.

Lexical learning revisited

As discussed in section 6.5 symbolic word acquisition approaches are likely to outperform our proposed lexical acquisition logic even under the assumption of a perfect syllable detector. This is mainly due to the integration of more word segmentation cues. Thus, a promising strategy to overcome this limitation, would be to integrate a metric segmentation mechanism into our framework. This would allow to incorporate the unique stress constraint outlined in chapter 2, which has been reported to greatly improve lexical acquisition in symbolic models. However, as to our best knowledge not even a developmentally plausible theoretical model for stress acquisition has been proposed yet. However, to assess the effect of metric cues it seems justifiable to weaken assumptions about developmental plausibility temporarily. As a consequence some assumptions about the stress model could be encoded into the model and the input signal. If metrical cues should turn out to actually improve lexical acquisition performance, stress cues may be added as additional speech feature dimension provided by specialized stress-highlighting signal-processing techniques.

The use of syllable transition probabilities equipped our lexical learning scheme with the ability to learn poly-syllabic words. However, it is not yet suited to cope with semantic based aggregations of bottom-up generated word candidates. For instance if $\mathcal{M}_W$ already contains the word `house` our proposed algorithm will not be able to learn `household` subsequently. But because there is evidence that children face the same problem, we consider our approach to be a first valid step in the direction of unsupervised lexical acquisition.

In this thesis we followed a straight-forward approach to model words as sequences of syllable symbols. This was to some extent motivated by computational simplicity. However, a natural next step would be to evaluate more elaborate word representations that are suited to overcome above mentioned insufficiencies. Especially it seems intriguing to model words as probability distributions over the space of syllable sequences. This we have investigated in section 7.3.3 for syllables, which failed (as expected) because of co-articulation effects. But as discussed above, words are assumed to be less prone to such effects, so we consider a probabilistic modeling of words to be a promising direction of further research. Furthermore, to validate our assumption about fewer co-articulation effects between syllables, words could be modeled as distinct acoustic model entities in a concatenative manner as implemented for syllables within this thesis.

In section 6.1.5 we proposed and discussed a plasticity scheme in which word acquisition was deferred compared to syllable learning. This was motivated by two arguments. First, we argued that the delay is necessary as words are learned from detected syllables that need to stabilize in advance. This also manifests in the delay term $d(t)$ to defer the learning of polysyllabic words. Second, we disregarded a fully time-decoupled delay scheme as used for phones and syllables, because we considered the number of syllables too numerous.

Now, after having investigated the process properties of our model in detail, we consider this to be an arguable assumption that should be further investigated. A delay of word learning could avoid the following divergence problem: until the syllable representation has not stabilized, the discriminative function of continuously adapted syllables may diverge from an initially associated word meaning. As the acoustic syllable structure of words may change over time, inferred words may no longer match the discriminative function of the words initially learned from the tutor.

Hence, a promising next step could be to delay lexical learning as long as the syllable representation has not been converged. This may ease the learning of words. Furthermore, this could result in an even more pronounced *lexical explosion effect* as residual learning, statistical learning, and mono-syllabic word learning would co-occur when lexical learning becomes initiated. Conceptually, the proposed lexical delay term $d(t)$ for polysyllabic words should be complemented with an additional regulatory pathway, that is directly coupled to the syllable speech coverage Γ. This would naturally shift the acquisition from syllables to words, as soon as the syllable representation has been stabilized. From a developmental point of view such a modification seems reasonable as lexical explosion takes place not before 24 months. Clearly, syllable segmentation and categorization abilities have been evolved almost completely at this point, so infants may rely on a similar fully decoupled acquisition scheme.

Evaluation of developmental processes

The approach that has been developed in this thesis attempts to provide a conceptually closed and complete computational model for speech structure acquisition. We have validated its implementation in detail on a sub-process level. To some extent we could show that our model is able to reflect some aspects of speech development. Our system seems to account for various findings from infant speech development: The plasticity of the phone representation vanishes after some time of habituation to a particular language, lexical explosion naturally emerges as a result of residual learning, and speech unit transition models facilitate the learning of less-granular speech representations. However, we could *not* prove its validity to full extent. This is due to several reasons.

First, we mainly employed semi-synthetic speech for evaluation. Only to a certain degree we have shown how the model performs in a more realistic setting using unconstrained read speech as input. As already discussed, we think this to be a valid and necessary approach given the complexity of the problem under consideration. However, to validate in greater detail whether and how our model may crack the code of language as infants do, it is clear that a more realistic type of input speech needs be evaluated. One possible next step could follow the argumentation of [Kit03, p. 9], who made the observation that many characteristics of child-directed speech, including the high frequency of phonological elisions and assimilations, are exactly the characteristics of adult-directed spontaneous speech. Thus instead of using rehearsed speech heard as input as pursued in this thesis, spontaneous speech corpora may be investigated.

Next, our embodied speech acquisition model relies only on speech as perceptual modality for speech structure learning. An integration of additional cues like gestures and mimics could allow to further constrain the learning task and may help to overcome some of the observed insufficiencies of our model. Whereas word acquisition is currently driven by speech coverage as sole criterion function to be maximized, more sophisticated task-models should be considered as driving forces of the emerging language abilities.

Finally and most importantly, the acquisition of cognitive functions as speech segmentation is a result of an interactive process with a parental tutor rather than a purely data-driven clustering process. However, only the proposed imitation learning interaction scheme partially accounts for this fact. To root learning more deeply in interaction, further evaluations should include the use of a (possibly embodied) tutoring agent that is able to provide a structurally and semantically rich input signal to the system. Clearly, this only defines a partial solution to the underlying problem, as an ultimate model for speech development would need to be validated by an actual human tutor in a long-lasting caregiver scenario comparable to the childhood of an infant.

Robustness and scalability

The model developed throughout this thesis has been built using the same pattern recognition methodology that is also employed for most ASR applications. As discussed in section 6.1, this streamlined the design and implementation process significantly and now paves the way for a future integration of standardized means to improve robustness and performance. This especially includes speaker-adaption methods which we neglected completely, or improved confidence computation schemes. Basically any new finding from ASR research is potentially also suited to improve the performance of the proposed model. This clearly distinguishes our approach from most other speech acquisition models as discussed in chapter 4 and 5.

Additionally, parts of the model may be replaced completely by more powerful solutions to further improve performance. For instance, the used Katz-backoff may be substituted by a more advanced discounting-scheme. Additionally, more efficient HMM-decoding techniques could help to increase the scalability of our model. Compared to the rather cluttered methodology landscape of competing models for speech modeling, we consider this to be an important conceptual and implementational advantage.

Form vs. function

An opposite direction for improvement would be to take not only results from developmental psychology into account, but to *learn from biology* itself. With the emergence of more advanced brain-imaging and studying techniques, auditory perception and categorical learning processes are underway to be understood by neuro-scientists. Hence, it seems intriguing to reflect such findings in computational models of these processes.

The approach to speech structure acquisition proposed by this thesis is exclusively functional only, as it neglects neural processing schemes as well as the organization of speech perception in the brain. Nevertheless, as a result of the modularity of the proposed system architecture, it seems reasonable to evaluate the use of more brain-like neuro-computational frameworks in the different subsystems. Especially we consider the regulatory control mechanisms to benefit from findings in neurobiology.

What comes after phones, syllables and words?

Although the different processing layers within our model are designed to learn the structure of speech on different levels of granularity, the conceptual structure of the different layers and their implementation are very similar. By following such a design paradigm our model has evolved naturally. Hence, an intuitive next step could be to extend the model with further layers. This we consider to be the most intriguing direction for future extension.

By retaining the conceptual idea of this thesis, a further layer would aim to model the next higher level of speech perception. This level we consider to represent the dialog model of the language. Such a *fourth layer to acquire dialog schemes* would be bootstrapped by observing interacting tutors. Its basic structure is depicted in figure 9.1. Its implementation would only gradually differ from that of the word layer. It would differ mainly in terms of the input features and the resulting "lexicon". The former would be the detected word utterances as obtained while the system attends tutor-tutor interactions. The lexicon would contain sequences of utterances. The transition

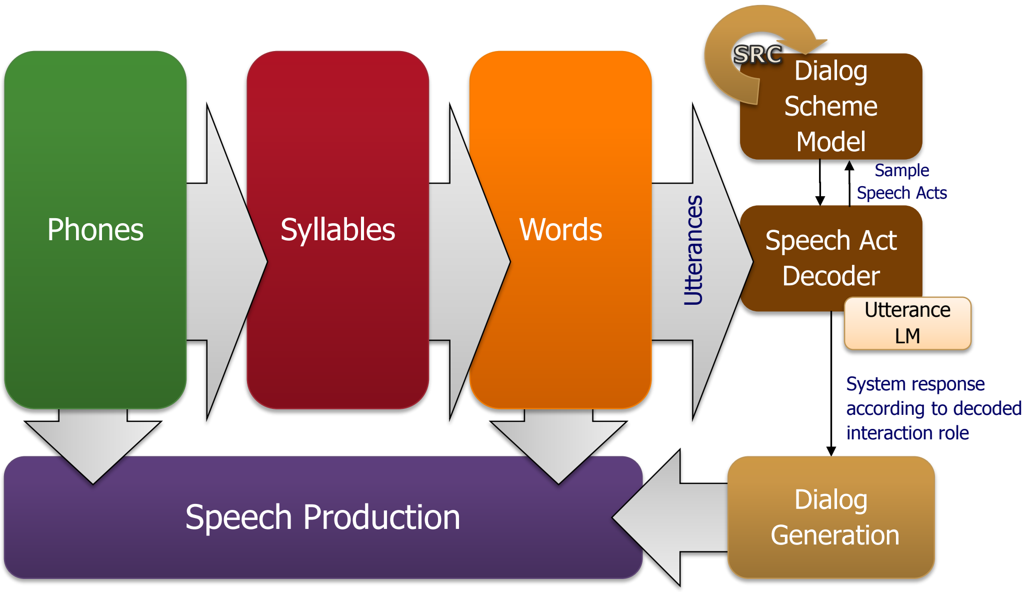


Figure 9.1: A fourth layer to acquire dialog schemes by observing two tutors in interaction. By applying generative principles similar to those used for speech production as presented in section 8.2, such an extended model could follow acquired dialog schemes in interaction with the tutor.

model of this additional layer would encode the probability that a particular utterance is followed by another particular one. Thus, the layer would aim to *condense speech acts into dialog schemes*. Bootstrapping of this layer would require our system to attend dialogs of two tutors similar to the experiments with the parrot ALEX [Pep98]. Promising results that skip phones, syllables, and word learning, but focus on a conceptually similar task of HMM-based *speech act detection* have been presented in [Rie99]. Clearly, to make such layer functional in an embodied agent, it would be necessary to enrich the system's input with scene representation features as outlined in section 8.1. Role-taking in such a system would naturally emerge by decoding system-directed utterances into speech acts.

Bibliography

[Abd03] S. A. Abdallah, M. D. Plumbley: *Probability as metadata event detection in music using ica as a conditional density model*, in *4th international symposium on independet component analysis and blind signal separation*, Nara, Japan, 2003, p. 233–238.

[Aka74] H. Akaike: *A new look at the statistical model identification*, *IEEE Transactions on Automatic Control*, volume 19, Nr. 6, 1974, p. 716–723.

[Akm01] A. Akmajian, R. A. Demers, A. K. Farmer, R. M. Harnish: *Linguistics - An Introduction to Language and Communication*, MIT Press, 2001.

[Ala99] A. Alani, M. Deriche: *A novel approach to speech segmentation using the wavelet transform*, in *Fifth International Symposium on Signal Processing and its Applications*, Brisbane, Australia, August 99.

[Asl98] R. N. Aslin, J. R. Saffran, E. L. Newport: *Computation of Conditional Probability Statistics by 8-month-old Infants*, *Psychological Science*, volume 9, Nr. 4, July 1998, p. 321–324.

[Au90] T. K. Au: *Children's use of information in word learning*, *Journal of Child Development*, volume 17, 1990, p. 393–416.

[Ave01] G. Aversano, A. Esposito, A. Esposito, M. Marinaro: *A new text-independent method for phoneme segmentation*, in *Proceedings of the 44th IEEE 2001 Midwest Symposium on Circuits and Systems*, 2001.

[Bal94] P. Baldi, Y. Chauvin: *Smooth On-Line Learning Algorithms for Hidden Markov Models*, *Neural Compuation*, volume 6, Nr. 2, 1994, p. 307–318.

[Bal03] D. H. Ballard, C. Yu: *A multimodal learning interface for word acquisition*, in *Proc. of ICASSP*, volume 5, 2003, p. 784–787.

[Bat08] A. Batliner, B. Schuller, S. Schaeffler, S. Steidl: *Mothers, adults, children, pets — Towards the acoustics of intimacy*, in *Proc. ICASSP*, 2008, p. 4497–4500.

[Baz00] I. Bazzi, J. Glass: *Modeling Out-of-vocabulary Words for Robust Speech Recognition*, in *Proc. ICSLP*, Beijing, 2000.

[Baz01] I. Bazzi, J. Glass: *Learning Units for Domain-Independent Out-of-Vocabulary Word Modelling*, in *Proc. Eurospeech*, 2001.

[Baz02] I. Bazzi: *Modeling Out-of-vocabulary Words for Robust Speech Recognition*, Dissertation, Massachusetts Institute of Technology, 2002.

[Bel05] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, M. B. Sandler: *A Tutorial on Onset Detection in Music Signals*, *IEEE Transactions on speech and audio processing*, volume 13, Nr. 5, September 2005, p. 1035–47.

[Ber04] N. Beringer: *Human language acquisition methods in a machine learning task*, IDSIA, 2004, nur als papier vorhanden.

[Bil04] J. A. Bilmes: *What HMMs can't do*, in *ATR Workshop "Beyond HMMs"*, 2004.

[Bis04] M. Bisani, H. Ney: *Bootstrap Estimates for Confidence Intervals in ASR Performance Evaluation*, in *Proc. ICASSP*, IEEE, May 2004, p. 409–412.

[Bis06] C. M. Bishop: *Pattern Recognition And Machine Learning*, Springer, 2006.

[Bis08] C. M. Bishop: *Pattern Recognition*, Microsoft Publishing, 2008.

[Bol08] B. Bolder, H. Brandl, M. Heracles, H. Janssen, I. Mikhailova, J. Schmüdderich, C. Goerick: *Expectation-driven Autonomous Learning and Interaction System*, in *IEEE-RAS Int. Conf. Humanoids – submitted*, 2008, submitted.

[Bor05] H. Bortfeld, J. L. Morgan, R. M. Golinkoff, K. Rathbun: *Mommy and Me*, *Psychological Science*, volume 16, Nr. 4, 2005, p. 298–304.

[Bra05] H. Brandl: *Sprechermodellierung auf geringen Trainingsstichproben*, Diplomarbeit, Ernst Moritz Arndt Universitaet Greifwald, 2005.

[Bra08] H. Brandl, F. Joublin, C. Goerick: *Towards unsupervised online word clustering*, in *Proc. ICASSP*, IEEE, 2008, p. 5073–76.

[Bre96] M. Brent, T. Cartwright: *Distributional regularity and phonotactic constraints are useful for segmentation*, *Cognition*, volume 61, 1996, p. 93–125.

[Bre01] M. Brent, J. Siskind: *The role of exposure to isolated words in early vocabulary development*, *Cognition*, volume 81, 2001, p. 33–44.

[Bro92] C. Browman, L. Goldstein: *Articulatory phonology: An overview*, *Phonetica*, volume 49, 1992, p. 155–180.

[Bur94] T. L. Burrows, M. Niranjan: *The use of recurrent neural networks for classification*, in *Proc. of the Workshop on Neural Networks for Signal Processing*, IEEE, Ermioni, Greece, September 1994, p. 117–125.

[Cai97] P. Cairns, R. Shillcock, N. Chater, J. Levy: *Bootstrapping word boundaries: a bottom-up corpus based approach to speech segmentation*, *Cognitive Psychology*, volume 33, 1997, p. 111–153.

[Car78] S. Carey: *The child as word learner*, in *Linguistic Theory and Psychological Reality*, MIT Press, 1978, p. 264–293.

[Cer08] C. Cerisara: *Automatic discovery of topics and acoustic morphemes from speech*, *Computer, Speech and Language*, 2008.

[Che05] L. Chen, H. Man: *Fast Schemes for Computing Similarities between. Gaussian HMMs and Their Applications*, *EURASIP Journal on Applied Signal Processing*, volume 13, 2005, p. 19841993.

[Cho65] N. Chomsky: *Aspects on the throry of syntax*, Dissertation, MIT, Cambridge, MA, 1965.

[Chr98] M. H. Christiansen, J. Allen, M. S. Seidenberg: *Learning to Segment Speech Using Multiple Cues: A Connectionist Model*, *Cognitive Science*, volume 14, 1998, p. 179–211.

[Cla03] S. Clark, J. R. Curran, M. Oshborne: *Bootstrapping POS taggers using unlabelled data*, in *Proc. of CONLL*, 2003.

[Cut87] A. Cutler, D. M. Carter: *The predominance of strong initial syllables in the English vocabulary.*, *Computer Speech and Language*, volume 2, 1987, p. 133–142.

[Cut88] A. Cutler, D. G. Norris: *The role of strong syllables in segmentation for lexical access*, *Journal of Experimental Psychology: Human Perception and Performance*, volume 14, 1988, p. 113–121.

[Cut96] A. Cutler: *Prosody and the word boundary problem*, Kap. 6, Erlbaum, Hillsdale, NJ, 1996, p. 87–99.

[Dav01] M. H. Davis: *Connectionist Modelling of Lexical Segmentation and Vocabulary Acquisition*, 2001.

[dBB99] B. de Boysson-Bardies: *How language comes to children*, MIT Press, 1999.

[Dom09] X. Domont: *Hierarchical spectro-temporal features for robust speech recogntion*, Dissertation, Technical University of Darmstadt, 2009.

[Dud00] R. O. Duda, P. E. Hart, D. G. Stork: *Pattern Classification*, Wiley-Interscience, 2nd. Ausg., October 2000.

[Duy02] P. Duygulu, K. Barnard, N. de Freitas, D. Forsyth: *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary*, in *Proc. Seventh European Conference on Computer Vision*, 2002, p. IV:97–112.

[Elm90] J. L. Elman: *Finding structure in time*, *Cognitive Science*, volume 14, 1990, p. 179–211.

[Fer98] A. Fernald, J. P. Pinto, D. Swingley, A. Weinberg, G. W. McRoberts: *Rapid gains in speed of verbal processing by infants in the second year*, *Psychological Science*, volume 9, 1998, p. 228–231.

[Fer02] S. Fernándex, A. Graves, H. Bunke, J. Schmidhuber: *An application of recurrent neural networks to discriminative keyword spotting*, in *Proc. 17th international conference on artificial neural networks*, 2002.

[FF03] L. Fei-Fei, R. Fergus, P. Perona: *A Bayesian Approach to Unsupervised One-Shot Learning of Object Categories*, in *Proc. of CCTV*, 2003.

[Fin03] G. A. Fink: *Mustererkennung mit Markovmodellen*, Leitfäden der Informatik, B. G. Teubner, 2003.

[Foo97] J. T. Foote, S. J. Young, G. J. F. Fones, K. S. Jones: *Unconstrained Keyword Spotting Using Phone Lattices with Application to Spoken Document Retrieval*, *Computer Speech and Language*, volume 11, 1997, p. 207–224.

[Foo01] J. Foote, S. Uchihashi: *The beat spectrum: A new approach to rhythm analysis*, in *Proc. of international conference on Multimedia and Expo*, IEEE, August 2001, p. 881–884.

[Fri94] B. Fritzke: *Fast learning with incremental RBF Networks*, *Neural Processing Letters*, volume 1, Nr. 1, 1994, p. 2–5.

[Gam05] T. Gambell, C. Yang: *Mechanisms and Constraints in Word Segmentation*, Yale University, June 2005.

[Gau94] J.-L. Gauvain, C.-H. Lee: *Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*, *IEEE Transactions SAP*, volume 2, 1994, p. 291–298.

[Gle81] S. M. Glenn, C. C. Cunningham, P. F. Joyce: *A study of auditory preference in non-handicapped infants and infants with downs syndrome*, *Child Development*, volume 52, 1981, p. 1303–7.

[Goe07] C. Goerick, B. Bolder, H. J. en, M. Gienger: *Towards Incremental Hierarchical Behavior Generation for Humanoids*, *IEEE-RAS International Conference on Humanoids*, 2007.

[Goe09] C. Goerick, J. Schmuedderich, B. Bolder, H. Janssen, M. Gienger, A. Bendig, M. Heckmann, T. Rodemann, H. B. X. Domont: *Interactive Online Multimodal Association for Internal Concept Building in Humanoids*, in *9th IEEE-RAS International Conference on Humanoid Robots*, 2009.

[Gol87] R. M. Golinkoff, K. Hirsh-Pasek, K. M. Cauley, L. Gordon: *The eyes have it: lexical and syntactic comprehension in a new paradigm*, *Journal of Child Language*, volume 14, 1987, p. 23–45.

[Gol05] S. Goldwater, M. Johnson: *Representational Bias in Unsupervised Learning of Syllable Structure*, in *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics, Ann Arbor, June 2005, p. 112–119.

[Gol06] K. Gold, B. Scassellati: *Audio Speech Segmentation Without Language-Specific Knowledge*, in *Cognitive Science*, Vancouver, 2006.

[Goo98] J. C. Goodman, L. McDonough, N. B. Brown: *The Role of Semantic Context and Memory In the Acquisition of Novel Nouns*, *Child Development*, volume 69, Nr. 5, 1998, p. 1330–1344.

[Gre97] S. Greenberg, B. E. D. Kingsbury: *The modulation spectrogram: In pursuit of an invariant representation of speech*, in *Proc. ICASSP*, IEEE, Munich, April 1997, p. 1647–1650.

[Gre98] S. Greenberg: *Speaking In Shorthand - A Syllable-Centric Perspective For Understanding Pronunciation Variation*, in *Proceedings of the ESCA Workshop on Modeling Pronounciation Variation for ASR*, 1998.

[Gro88] S. Grossberg: *Nonlinear neural networks: principles, mechanisms, and architectures*, *Neural Networks*, volume 1, 1988, p. 17–61.

[Har95] B. Hart, T. R. Risley: *Meaningful differences in the everyday experience in young american children*, Paul H Brookes, 1. Ausg., 1995.

[Hec09] M. Heckmann, H. Brandl, J. Schmuedderich, X. Domont, B. Bolder, I. Mikhailova, H. Janssen, F. Joublin, C. Goerick: *Teaching Asimo: Audio-Visual Association Learning in Headset-Free Interaction*, in *Proc. ROMAN*, 2009.

[Hei87] T. H. Heibeck, E. M. Markmann: *Word learning in children: An examination of fast mapping*, *Child Development*, volume 58, 1987, p. 1021–1034.

[Hic09] G. Hickok: *What is speech perception?*, http://talkingbrains.blogspot.com/2009/03/what0is-speech-perception.html, March 2009, blog entry.

[Hoc00] S. Hochreiter, J. Schmidhuber: *Long Short-Term Memory*, *Neural Computation*, volume 9, Nr. 8, 2000, p. 1735–1780.

[Hsi99] C.-T. Hsieh, M.-C. Su, E. LAI, C.-H. Hsu: *A Segmentation Method for Continuous Speech Utilizing Hybrid Neuro-Fuzzy Network*, *Journal of Information Sciences and Engineering*, volume 15, 1999, p. 615–628.

[HT03] D. Hakkani-Tür, G. Riccardi: *Active And Unsupervised Learning For Automatic Speech Recognition*, in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2003.

[Hua01] X. Huang, A. Aceero, H.-W. Hon: *Spoken Language Processing*, Prentice Hall PTR, 2001.

[Hun00] A. Hunt: *JSpeech Grammar Format*, http://www.w3.org/TR/jsgf/, June 2000.

[Iwa03] N. Iwahashi: *Language acquisition through a human-robot interface by combining speech, visual, and behavioral information*, *Information Sciences*, volume 156, 2003, p. 109–121.

[Iwa04] N. Iwahashi: *Active and unsupervised learning of spoken words through a multimodal interface*, in *Proc. 13th IEEE Workshop Robot and Human Interactive Communication*, 2004, p. 437–442.

[Iwa06] N. Iwahashi: *Robots that Learn Language: Developmental Approach to Human-Machine Conversations*, in P. Vogt, Y. Sugita, E. Tuci, C. Nehaniv (Hrsg.): *Symbol Grounding and Beyond - EELC*, 2006, p. 143–167.

[Jel97] F. Jelinek: *Statistical Methods for Speech Recognition*, MIT Press, 1997.

[Jit98] N. Jittiwarangkul, S. Jitapunkul, S. Luksaneeyanawin, V. Ahkuputra, C. Wutiwiwatchai: *Thai Syllable Segmentation for Connected Speech Based on Energy*, in *Proc. Asia-Pacific conference on circuits and systems*, IEEE, November 1998, p. 169–172.

[Jua85] B.-H. Juang, L. R. Rabiner: *A Probabilistic Distance Measure for Hidden Markov Models*, *AT&T Technical Journal*, volume 64, 1985, p. 391–408.

[Jun00] J. Junkawitsch: *Detektion von Schlsselwortern in fliessender Sprache*, Dissertation, Technical University of Munich, 2000.

[Jun03] A. Juneja, C. Espy-Wilson: *Speech Segmentation Using Probabilistic Phonetic Feature Hierarchy and Support Vector Machines*, in *Proc. of the international joint conference on neural networks*, IEEE, 2003, p. 675–679.

[Jus93a] P. Jusczyk, A. Cutler, N. Redanz: *Preference for the predominant stress patterns of English worlds*, *Child Development*, volume 64, 1993, p. 675–687.

[Jus93b] P. W. Jusczyk, A. D. Friederici, J. M. Wessels, V. Y. Svenkerud, A. M. Jusczyk: *Infants sensitivity to the sound patterns of native language words*, *Journal of Memory and Language*, volume 32, 1993, p. 402–420.

[Jus95] P. W. Jusczyk, R. N. Aslin: *Infants' detection of the sound patterns of words in fluent speech*, *Cognitive Psychology*, volume 29, 1995, p. 1–23.

[Jus97a] P. W. Jusczyk: *The discovery of spoken language*, MIT Press, Cambridge, MA, 1997.

[Jus97b] P. W. Jusczyk, E. A. Hohne: *Infants memory for spoken words*, *Science*, volume 277, 1997, p. 1984–1985.

[Jus99a] P. Jusczyk, E. A. Hohne, A. Baumann: *Infants sensitivity to allophonic cues for word segmentation*, *Perception and Psychophysics*, volume 61, 1999, p. 1465–1476.

[Jus99b] P. W. Jusczyk: *How infants begin to extract words from speech*, *Trends in Cognitive Sciences*, volume 3, Nr. 9, September 1999, p. 323–328.

[Jus99c] P. W. Jusczyk, D. M. Houston, M. Newsome: *The beginnings of word segmentation in English-learning infants*, *Cognitive Psychology*, volume 39, 1999, p. 159–207.

[Kah76] D. Kahn: *Syllable-based Generalizations in English Phonology*, Dissertation, University of Massachusetts, 1976.

[Kam00] S. O. Kamppari, T. J. Hazen: *Word and phone level acoustic confidence scoring*, in *Proc. ICASSP*, volume 3, Istanbul, 2000, p. 1799–1802.

[Kat74] N. Katz, E. Baker, J. MacNamara: *What's in a name? A study of how children learn common an proper names*, *Child Development*, volume 45, 1974, p. 469–473.

[Kat87] S. M. Katz: *Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer*, *Acoustics, Speech and Audio Processing, IEEE Transactions on*, volume 3, 1987, p. 400–401.

[Kem99] T. Kemp, A. Waibel: *Unsupervised Training Of A Speech Recognizer: Recent Experiments*, in *Proc. Eurospeech*, 1999, p. 2725–2728.

[Ket06] H. Ketabdar, J. Vepa, S. Bengio, H. Bourlard: *Posterior Based Keyword Spotting with A Priori Thresholds*, in *Proc. ICSLP*, Pittsburgh, 2006.

[Ket07] H. Ketabdar, H. Hermansky: *Detection of Out-of-Vocabulary Words in Posterior Based ASR*, in *Proc. Interspeech*, 2007.

[Kit03] C. Kit: *How Does Lexical Acquisition Begin? A Cognitive Perspective*, *Cognitive Science*, volume 1, 2003, p. 1–50.

[Kla99] A. Klapuri: *Sound onset detection by applying psychoacoustic knowledge*, in *Proc. ICASSP*, Phoenix, Arizona, 1999.

[Koh88] T. Kohonen: *The "neural" phonetic typewriter*, *Computer*, volume 21, Nr. 3, 1988, p. 11–22.

[Koh89] T. Kohonen: *Self-Organizaion and Associative Memory*, Springer-Verlag, Berlin, Germany, 3rd. Ausg., 1989.

[Kri93] V. Krishnamurthy, J. B. Moore: *On-Line Estimation of Hidden Markov Model Parameters Based on the Kullback-Leibler Information Measure*, *IEEE Transactions on Signal Processing*, volume 41, Nr. 8, August 1993, p. 2557–2573.

[Lad93] P. Ladefoged: *A course in phonetics*, TX: Harcourt Brace, 1993.

[Lam01] L. Lamel, J.-L. Gauvain, G. Adda: *Investigating Lightly Supervised Acoustic Model Training*, *ICASSP*, 2001.

[Lam02a] L. Lamel, J.-L. Gauvain, G. Adda: *Lightly Supervised and Unsupervised Acoustic Model Training*, *Computer, Speech and Language*, volume 16, 2002, p. 115–129.

[Lam02b] L. Lamel, J. luc Gauvain, G. Adda: *Unsupervised Acoustic Model Training*, in *Proc. of ICASSP*, volume 1, Orlando, May 2002, p. 877–880.

[Leg95] C. Leggetter, P. C. Woodland: *Maximum likelihood linear regression for speaker adaption of continuous density hidden Markov models*, *Computer, Speech and Language*, volume 9, Juni 1995, p. 171–185.

[Leo84] R. G. Leonard: *A database for speaker-independent digit recognition*, *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP*, volume 9, 1984, p. 328–331.

[Leo93] R. G. Leonard, G. Doddington: *TIDIGITS*, Linguistic Data Consortium, Philadelphia, 1993.

[Lev05] S. Levinson, K. Squire, R.-S. Lin, M. McClain: *Automatic Language Acquisition by an Autonomous Robot*, in *AAAI Spring Symposium on Developmental Robotics*, 2005.

[Ma02] B. Ma, Q. Huo: *A Comparative Study of Several Incremental Adaptation Algorithms for Speaker Adaptation*, in *Proc. ISCSLP*, Taipei, August 2002, p. 347–350.

[Mac95] B. MacWhinney: *The CHILDES project: Tools for analyzing talk*, Erlbaum, Hillsdale, New York, 2nd. Ausg., 1995.

[Mar95] C. D. Marcken: *Acquiring a Lexicon from Unsegmented Speech*, in *Meeting of the Association for Computational Linguistics*, 1995, p. 311–313.

[Mar03] M. Markou, S. Singh: *Novelty Detection: A Review - Part 2: Neural Approaches*, *Signal Processing*, volume 83, 2003, p. 2499–2521.

[Mar07] K. Markov, S. Nakamura: *Never-Ending Learning with Dynamic Hidden Markov Network*, in *Proc. Interspeech*, August 2007, p. 1437–1440.

[Meh88a] J. Mehler, P. W. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, C. Amiel-Tison: *A precursor of language acquisition in young infants*, *Cognition*, volume 29, 1988, p. 143–178.

[Meh88b] J. Mehler, P. W. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, C. Amiel-Tison: *A precursor of language acquisition in young infants*, *Cognition*, volume 29, 1988, p. 143–78.

[Mei99] H. Meinedo, J. P. Neto, L. B. Almeida: *Syllable Onset Detection Applied To The Portuguese Language*, in *Proceedings of Eurospeech*, 1999.

[Mei02] M. Meila: *Comparing Clusterings*, University of Washington, 2002.

[Mih04] R. Mihalcea: *Co-training and self-training for word sense disambiguation*, in *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, 2004.

[Mik08] I. Mikhailova, M. Heracles, B. Bolder, H. Janssen, H. Brandl, J. Schmüdderich, C. Goerick: *Coupling of mental concepts to a reactive system: incremental approach in system design*, in *Submitted to the Eighth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 2008.

[Mpo08] I. Mporas, T. Ganchev, N. Fakotakis: *A hybrid architecture for automatic segmentation of speech waveforms*, in *Proc. ICASSP*, 2008, p. 4457–60.

[Mur04] H. A. Murthy, T. Nagarajan, N. Hemalatha: *Automatic Segmentation and Labeling of Continuous Speech Without Bootstrapping*, in *Proc. of EUSIPCO*, 2004, Poster-presentation.

[Nag03] T. Nagarajan, H. A. Murthy, R. M. Hegde: *Segmentation of speech into syllable-like units*, in *Proc. Eurospeech*, Geneva, 2003, p. 2893–96.

[Noe91] A. Noetzel: *Robust Syllable Segmentation Of Continuous Speech Using Neural Networks*, in *Electro International Conference Record*, IEEE, April 1991, p. 580–585.

[Nor94] D. Norris: *Shortlist: a connectionist model of continuous speech recognition*, *Cognition*, volume 52, 1994, p. 189–234.

[Pep98] I. Pepperberg: *Talking with Alex: Logic and speech in parrots*, in *Exploring Intelligence*, Scientific American, 1998, p. 35–38.

[Pet83] A. Peters: *The units of language acquisition*, MIT Press, MA, 1983.

[Pet96] B. Petek, O. Andersen, P. Dalsgaard: *On the robust automatic segmentation of spontaneous speech*, in *Proc. ICSLP*, 1996, p. 913–116.

[Pfi96] H. R. Pfitzinger, S. Burger, S. Heid: *Syllable detection in read and spontaneous speech*, in *Proc. ICSLP*, 1996, p. 1261–1264.

[Pin89] S. Pinker: *Learnability and cognition*, MIT Press, Cambridge, MA, 1989.

[Pl01] T. Pltz: *Online-Adaption statistischer Spracherkennungssysteme*, Dissertation, Universit Bielefeld, Technische Fakultt, Angewandte Informatik, September 2001.

[Pla92] B. Plannerer, G. Ruske: *Recognition of demisyllable based units using semicontinuous hidden Markov models*, in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, 23-26 March 1992, p. 581–584vol.1.

[Pwi05] M. Pwint, F. Sattar: *A segmentation method for noisy speech using genetic algorithm*, in *Proc. ICASSP*, 2005.

[Qia08] Y. Qiao, N. Shimomura, N. Minematsu: *Unsupervised optimal phoneme segmentation: Objectives, Algorithms and comparisons*, in *Proc. ICASSP*, 2008, p. 3989–92.

[Qui60] W. W. O. Quine: *Word and Object*, MIT Press, Cambridge, MA, 1960.

[Rab89] L. R. Rabiner: *A Tutorial on Hidden Markov Models and selected applications in speech recognition*, in *Proc. of IEEE*, volume 77, 1989, p. 257–286.

[Rie99] K. Ries: *Hmm and neural network based speech act detection*, in *Proc. ICASSP*, IEEE, Arizona, USA, March 1999.

[Rig98] G. Rigoll: *Hybrid speech recognition systems - A real alternative to traditional approaches*, in *Survey Lecture, Proc. International Workshop Speech and Computer (SPECOM'98)*, 1998.

[Roh04] K. J. Rohlfing, J. Fritsch, B. Wrede: *Learning to manipulate objects: A quantitative evaluation of Motionese*, in *Proceedings of the Third International Conference on Development and Learning (ICDL 2004)*, 2004.

[Rom07] Romei: *Occiptial TMS has opposing effects on Auditory Stimulus detection*, *Journal of Neuroscience*, 2007.

[Roy99] D. Roy: *Learning Words from Sights and Sounds: A Computational Model*, Dissertation, MIT, 1999.

[Roy00] D. Roy: *A Computational Model of Word Learning from Multimodal Sensory Input*, in *International conference of Cognitive Modeling*, 2000.

[Roy06] D. Roy, R. Patel, P. DeCamp, R. Kubat, M. Fleischmann: *The Human Speechome Project*, in *Annual Conference of the Cognitive Science Society*, July 2006.

[Rus81] G. Ruske, T. Schotola: *The Efficiency of Demisyllable Segmentation in the Recognition of Spoken Words*, in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta,, 1981, p. 971–974.

[Saf96] J. R. Saffran, R. N. Aslin, E. L. Newport: *Statistical language learning by 8 month old infants*, *Science*, volume 274, 1996, p. 1926–1928.

[San98] A. Sankar: *Experiments with a Gaussian Merging-Splitting Algorithm for HMM Training for Speech Recognition*, in *Proceedings of DARPA Speech Recognition Workshop*, Lansdowne, VA, February 1998.

[Sar04] G. Sarada, N. Hemalatha, T. Nagarajan, H. A. Murthy: *Automatic Transcription of Continuous Speech using Unsupervised and Incremental Training*, 2004, Poster-Presentation at InterSpeech 2004.

[Sat03] N. Satravaha, P. Klinkhachom, N. Lass: *Tone classification of syllable-segmented Thai speech based on multilayer perceptron*, in *Proc. of the 35th south-eastern symposium on system theory*, March 2003, p. 392–396.

[Sch05] C. Schrumpf, M. Larson, S. Eickeler: *Syllable-based language model in speech recognition for English spoken document retrieval*, in *Proc. of 7th International workshop of the EU network of excellence DELOS on audio-visual conent and information visualization in digital libraries*, 2005.

[Sha99] L. Shastri, S.Chang, S. Greenberg: *Syllable Detection And Segmentation Using Temporal Flow Neural Networks*, in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999.

[Sha07] F. Sha, L. K. Saul: *Large margin hidden markov models for automatic speech recognition*, in *Advances in Neural Information Processing Systems*, volume 19, MIT Press, 2007.

[Shi97] M. L. Shire: *Syllable onset detection from acoustics*, Diplomarbeit, University of California, Berkely, May 1997.

[Sim89] J. Simpsion: *Oxfored English Dictionary*, Oxford Univerisity Press, 2nd. Ausg., 1989.

[Squ05] K. M. Squire: *HMM-Based Semantic Learning for a Mobile Robot*, *IEEE Transactions on evolutionary computing*, volume X, Nr. X, 2005, p. 1–14.

[Str35] J. R. Stroop: *Studies of interference in serial verbal reactions*, *Jounral of Experimental Psychology*, volume 18, 1935, p. 643–662.

[Swi05] D. Swingley: *Statistical clustering and the contents of the infant vocabulary*, *Cognitive Psychology*, volume 50, 2005, p. 86–132.

[Tah01] S. M. Tahir, A. Z. Shaámeri, S. H. S. Salleh: *Time-varying autoregressive modeling approach for speech segmentation*, in *International Symposium on Signal Processing and its Applications*, Kuala Lumpr, Malaysia, August 2001, p. 715–717.

[Tao02] J. Tao, H. U. Hain: *Automatic speech segmentation for Chinese speech database based on HMM*, in *Proceedings of TENCON' 02*, volume 1, 2002, p. 481–484.

[Tay88] M. Taylor, S. A. Gelman: *Adjectives and nouns: Children strategies for learning new words*, *Child Development*, volume 59, 1988, p. 411–419.

[The03] S. Theodoridis, K. Koutroumbas: *Pattern Recognition*, Elsevier Academic Press, London, 2. Ausg., 2003.

[Thi03] E. Thiessen, J. R. Saffran: *When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants*, *Developmental Psychology*, volume 39, 2003, p. 706–716.

[Tho81] D. G. Thomas, J. J. Campos, D. W. Shucard, D. S. Ramsay, S. J.: *Semantic comprehension in infancy: A signal detection analysis*, *Child Development*, volume 52, 1981, p. 798–903.

[Tol04] D. T. Toledano, L. A. H. Gmez: *HMMs for Automatic Phonetic Segmentation*, 2004, SP3 Annotation Tools: From Speech Segments to Dialouges.

[Var06] Various: *Poems Every Child Should Know*, Kessinger Publishing, LLC, 2006.

[Var09] Various: *LibriVox - Acoustical liberation of books in the public domain*, http://librivox.org/, 2009.

[Vaz08] M. Vaz, H. Brandl, F. Joublin, C. Goerick: *Linking perception and production: system learns a correspondence between its own voice and the tutor's*, in *Workshop about Speech and Face to Face communication*, 2008.

[Vaz09a] M. Vaz: *A developmentally inspired computational framework for embodied speech imitation*, Dissertation, Universidade do Minho, 2009.

[Vaz09b] M. Vaz, H. Brandl, F. Joublin, C. Goerick: *Learning from a tutor: Embodied speech acquisition and imitation learning*, in *Proc. IEEE 8th International Conference on Development and Learning*, Shanghai, China, 05.07.2009 2009.

[Vaz09c] M. Vaz, H. Brandl, F. Joublin, C. Goerick: *Speech imitation with a child's voice: addressing the correspondence problem*, in *Proc. SPECOM'2009 13-th International Conference on Speech and Computer*, St Petersburg, Russia, June 2009, p. 289–294.

[vH91] J. van Hemert: *Automatic Segmentation of Speech*, *IEEE Trans. on Signal Processing*, volume 39, Nr. 4, 1991, p. 1008–1012.

[Wai89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang: *Phoneme Recognition Using Time-Delay Neural Networks*, *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 37, Nr. 3, 1989, p. 328–339.

[Wal93] A. Walley: *The role of vocabulary development in children's spoken word recognition and segmentation ability*, *Developmental Review*, volume 13, 1993, p. 286–350.

[Wal04] W. Walker, P. Lamere, P. Kwok: *Sphinx-4: A Flexible Open Source Framework for Speech Recognition*, 2004.

[Wan03] D. Wang, L. Lu, H.-J. Zhang: *Speech segmentation without speech recognition*, in *Proc. ICASSP*, 2003.

[Wat04] S. Watanabe, A. Sako, A. Nakamura: *Automatic Determination Of Acoustic Model Topology Using Variational Bayesian Estimation and Clustering*, *ICASSP*, volume 1, 2004, p. 813–816.

[Wax96] S. Waxman, D. Markow: *Words as an invitation to form categories: Evidence from 12- to 13-month-olds*, *Cognitive Psychology*, volume 29, 1996, p. 257–302.

[Wei95] M. Weintraub: *LVCSR log-likelihood ratio scoring for keyword spotting*, in *Proc. of ICASSP*, volume 1, 1995, p. 297–300.

[Wes01] F. Wessel, H. Ney: *Unsupervised Training Of Acoustic Models For Large Vocabulary Continuous Speech Recognition*, in *Automatic Speech Recognition and Understanding Workshop*, 2001.

[Wu97] S.-L. Wu, M. L. Shire, S. Greenberg, N. Morgan: *Integrating syllable boundary information into speech recognition*, in *Proc. ICASSP*, volume 2, Munich, 1997, p. 987–990.

[Wu98] S.-L. Wu: *Incorporating information from syllable-length time scales into automatic speech recogntion*, Dissertation, University of California, 1998.

[You06] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw: *The HTK Book for HTK V3.4*, Cambridge University Press, Cambridge, UK, 2006.

[Yu08] C. Yu, L. B. Smith, A. F. Pereira: *Embodied solution: The world from a toddler's point of view*, in *Proceedings of IEEE 7th International Conference in Development and Learning*, IEEE, 2008.