Res on Lang and Comput (2010) 8:133–168
DOI 10.1007/s11168-011-9075-4

A Computational Model of Unsupervised Speech Segmentation for Correspondence Learning

Daniel Duran · Hinrich Schütze · Bernd Möbius · Michael Walsh

Published online: 23 April 2011

Abstract In this paper, we develop a new conceptual framework for an important problem in language acquisition, the correspondence problem: the fact that a given utterance has different manifestations in the speech and articulation of different speakers and that the correspondence of these manifestations is difficult to learn. We put forward the Correspondence-by-Segmentation Hypothesis, which states that correspondence is primarily learned by first segmenting speech in an unsupervised manner and then mapping the acoustics of different speakers onto each other. We show that a rudimentary segmentation of speech can be learned in an unsupervised fashion. We then demonstrate that, using the previously learned segmentation, different instances of a word can be mapped onto each other with high accuracy when trained on utterance-label pairs for a small set of words.

Keywords Language acquisition · Speech · Perception · Production · Correspondence learning · Segmentation

1 The Correspondence Problem

A fundamental component of linguistic competence is to recognize different percepts as corresponding to the same underlying phone. For example, when I produce the English phone [ð], (i) I observe myself directing the tongue to move towards the teeth (motoric), (ii) I hear a particular type of frication (auditory), and (iii) I feel my tongue touching the teeth (haptic). When my interlocutor produces [ð], (iv) I hear a particular type of frication (auditory input that can be quite different from the input I receive

D. Duran · H. Schütze (✉) · B. Möbius · M. Walsh
Institute for Natural Language Processing, University of Stuttgart, Azenbergstr 12, 70174 Stuttgart, Germany
e-mail: hinrich+rolc2011@gmail.com

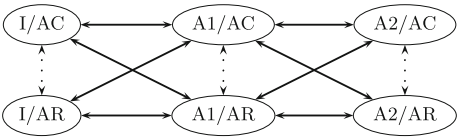


Fig. 1 Possible combinations of input sources from the child (I = infant), caregiver 1 (A1 = adult 1), and caregiver 2 (A2 = adult 2). In each case, both acoustics (AC) and articulation (AR) are considered. The links are to be interpreted from the perspective of the child. For example, articulation in I/AR is stored as motoric and sensory percepts; articulation is stored as visual percepts in A1/AR and A2/AR

when I produce [ð] myself), and (v) I see movements of some of my interlocutor's articulators (lips and tongue). Mature speakers of a language have learned correspondences between all of these different manifestations of a phone. They know which motor commands to use to produce it; they correct motor commands if they "feel" that the articulators did not go where they were supposed to go; they correct speech if based on auditory feedback from their own production they realize that they did not pronounce something correctly; they understand their interlocutors based on hearing them; and they can (to a limited extent) lip read (Rosenblum 2008).

Figure 1 shows the major correspondences that have to be learned in language acquisition, focusing on acoustics and articulation. We use I, A1, A2, AC, and AR to refer to infant, adult 1, adult 2, acoustics and articulation, respectively. Each link in the figure represents a type of correspondence the child has to learn. For example, I/AC-A1/AC is the correspondence between the acoustic signal produced by the child for a particular phone and the acoustic signal produced by the adult for that same phone.

Three of these links, I/AC-I/AR, A1/AC-A1/AR and A2/AC-A2/AR (dotted in the figure), are *synchronous:* The child perceives the acoustics and articulation of a phone she produces at (almost) exactly the same time. The same holds for the adult equivalents A1/AC-A1/AR and A2/AC-A2/AR. Learning that opening the lips while there is air pressure in the mouth causes a particular type of sound to occur ([b] or [p]) is similar to many other basic facts about the world the child has to learn. For example, the child needs to learn that, when moving the left hand a certain way, it will suddenly appear in front of her eyes. This is one of many nontrivial tasks the child must learn in addition to the other correspondences in Fig. 1.

The other links (those except for I/AC-I/AR, A1/AC-A1/AR, and A2/AC-A2/AR) are harder because they are asynchronous. Consider the case A1/AC-A2/AC. The child is unlikely to hear a production of [ð] of adult 1 and a production of [ð] of adult 2 at the same time—and if she does it is likely to prove confusing rather than helpful in learning the correspondence. Instead, the two instances usually occur separated in time.

A further, and even more difficult case, is the link I/AC-A1/AC. There are significant differences between what the child perceives (i) while producing a sequence of sounds (e.g., "bababa") and (ii) while listening to an adult producing the same sequence of sounds. This is due to the fact that the child's articulatory apparatus is different from the adult's and due to the fact that the child's articulations have not reached adult competence yet. Perhaps most important is the difference caused by resonances and

sound waves within the child's body versus the different acoustic forces acting upon a sound wave that travels from the adult to the child through the air.

Because of these significant differences a simple correlation analysis operating directly on the speech signal is unlikely to be the basis for learning correspondence; and we know of no work that has attempted to show that this is possible. However, it seems plausible that a simple correlation mechanism operating on a *higher* level of representation and a small number of perceptual categories (such as phones) would be able to learn the correspondence between child and adult language.

Consider two productions of "bababa", one production i by the infant with length 580 ms and one production a by the adult with length 610 ms. A low-level signal representation of i and a into, say, 100 acoustic events would yield hundreds if not thousands of possible alignments. (There would be far fewer than 10,000 ($100 \cdot 100$) because the alignment has to be monotonic: for acoustic events $a_1 < a_2$ and $i_1 < i_2$, where $<$ indicates temporal precedence, the crossing alignment (a_1, i_2), (a_2, i_1) is not admissible.)

In contrast, if the child is able to analyze i as [silence_i – plosion_i – vowel_i – silence_i – plosion_i – vowel_i – silence_i – plosion_i – vowel_i] and a as [silence_a – plosion_a – vowel_a – silence_a – plosion_a – vowel_a – silence_a – plosion_a – vowel_a], then the number of alignments (which depends on the exact formalization of monotonicity) is much smaller. Of course, the labels are misleading—even if the infant is able to do this analysis, she will initially not be able to recognize vowel_i and vowel_a as instances of the same phone as infant vowels and adult vowels are acoustically different.

Nevertheless, there is only one possible solution if the additional constraint is imposed that infant category i' and adult category a' have to be consistently aligned. In this example, this would mean that the two silence categories, the two plosion categories and the two vowel categories are aligned. Crucially, this would be possible without any assumption that the infant can recognize its own vowels and adult vowels as instances of the same category.

It is unclear how correspondence could be established without some form of segmentation of the kind we have just described; and no alternative realistic computational model of correspondence learning currently exists that would explain how infants learn correspondence between phones. We will show below that segmentation into high-level acoustic events is possible in an unsupervised fashion. Thus, all the information necessary for performing this analysis is, in principle, available to the infant. Our fundamental assumption in this paper is that children can learn segmentation in an unsupervised fashion and that this segmentation is the basis for correspondence learning:

Correspondence-by-Segmentation Hypothesis (CSH)

The infant learns correspondence by (i) segmenting the speech of person 1 (infant, adult 1, or adult 2) into a sequence of high-level acoustic categories; (ii) segmenting the speech of a different person (infant, adult 1, or adult 2) into a sequence of (different) high-level acoustic categories; and (iii) aligning these two segmentations for productions that are recognized as being the same or similar.

This hypothesis is in keeping with much research in phonetics and child language acquisition, e.g., Kuhl (1987, p.364) writes:

> We know, then, that infants' representation of these syllables [/ma/, /mi/ and /mu/] allows them to break the syllables down into some kind of 'parts'-ones that allow them to detect similarity at the beginnings of the syllables in spite of differences at the ends of syllables. At the very least, then, this ability must rely on a representation of units that allows *portions of syllables* to be isolated and compared across syllables.

We do not address the question of how the child recognizes certain pairs of sound sequences as referring to the same object. One simple case is that the adult repeats what the child was saying (e.g., the adult saying a_1 = "bababa" in response to the child saying i_1 = "bababa") and that the child then assumes that a_1 is a repetition of i_1. Another simple case is that adult 1 and adult 2 take turns referring to the same object repeatedly.

The CSH as stated above hypothesizes that acoustics are the primary means of learning correspondence. Modified versions for links involving articulation would assume an alignment of high-level acoustic events with articulatory events (I/AR-A1/AC) or visual events (I/AC-A1/AR).

The link I/AR-A1/AR does not involve acoustics, but we know of no claims in the literature that it plays an important role in correspondence learning. Although there is evidence that neonates "can imitate both facial and manual gestures" (Meltzoff and Moore 1977), it seems hard to imagine how speech segmentation could be achieved based primarily, or, exclusively, on this correspondence link. The general importance of visual information in speech perception (A1/AR) is, however, investigated in the context of speech as a multimodal or amodal phenomenon (Kuhl and Meltzoff 1982; Massaro 2004; Fowler 2004; Rosenblum 2008).

Thus, the links I/AR-A1/AC and I/AC-A1/AR require part of the acoustic segmentation posited by the CSH. So even if our hypothesis that the links I/AC-A1/AC and A1/AC-A2/AC are the primary drivers of correspondence learning is not correct, acoustic segmentation might still be required for learning correspondence.

The computational model we will present below is not intended to model the details of the actual computational mechanisms that a child would be able to employ. For example, the model is a batch processing model whereas human computing is online and highly parallel. Our goal is to show that segmentation and correspondence can in principle be learned based on the information that is available to the child. We will not address the details of how the human cognitive system avails itself of this information.

In summary, correspondence, the ability to recognize percepts from different modalities as instances of the same phone, is one of the fundamental components of language competence. There is currently no computational model of how correspondence could be learned on realistic speech data. In this paper, we present such a model, based on the Correspondence-by-Segmentation Hypothesis, which states that correspondence is learned by performing parallel segmentations of the speech of two different persons (e.g., adult 1 and adult 2) and then aligning the segments. Given that particular terms,

such as *segment*, can mean different things to different people, we next specify our terminology before proceeding with the remainder of the article.

Segments and Phones

The term *segment* can refer to units on various levels of linguistic abstraction: phones, phonemes, syllables, words, and utterances. Generally, we will refer to intervals as segments if they were *learned by our computational model*. In phonetics, *segment* also has a more specific meaning: "any discrete unit that can be identified, either physically or auditorily, in the stream of speech" (Crystal 2003). We will avoid this more specific meaning of segment and use the term *phone* instead.

We define *phone acquisition* as the problem of unsupervised learning of phones from the speech signal. The phones learned in this fashion are referred to as *induced phones*, and linguistically defined and annotated phones as *labeled phones*.

Supervised, Semi-supervised and Unsupervised Learning

A classification problem is *supervised* if labeled training examples are provided to the learning algorithm. For example, learning of segmentation is supervised if a training set with gold standard segmentation boundaries is used.

For the purposes of this paper, we define a learning problem as *unsupervised* if the learning algorithms does not have access to any labels, that is, no access to any classification decisions made by annotators on training data. We treat phone acquisition as an unsupervised learning problem. The learning algorithm operates without any human-generated labels.

We define a learning problem as *semi-supervised* if the learning algorithm has access to human labels, but the task of the classifier to be learned is different from the human labeling task. We treat the correspondence problem as a semi-supervised learning problem in this paper. The learning algorithm has access to human labels, but labels of utterance-length *intervals of speech*, not labels of *pairs of corresponding phones*. For example, we assume that the child is able to recognize that two different intervals of speech both refer to an elephant; in this case, "elephant" is the human label assigned to these two intervals and these labels are then the input to the learning algorithm.

The difference between semi-supervised and supervised learning is subtle in that the same experiment can be viewed as either semi-supervised or supervised, depending on which aspect of the experiment is in focus. Our evaluation of the success of correspondence learning is indirect: we do not provide direct measures of the quality of the acquired correspondences. This is difficult to do because the induced phones whose correspondence is at issue are not traditional linguistic categories. Thus, it is difficult to inspect and understand them and even more difficult to say with any certainty how well two of them correspond to each other. For this reason, we evaluate the success of correspondence learning on a word recognition task. This word recognition task is clearly a supervised learning task: manually labeled training examples of the exact task to be learned are provided to the learning algorithm. Thus, if viewed as a

word recognition task, the experiment in Sect. 8 is supervised learning; if viewed as a phone correspondence learning task, then it is semi-supervised.

It is important to stress that even a completely unsupervised learning algorithm is not a discovery procedure that magically discovers structure in data without the human subjectivity that is inherent in labels. For example, we represent the acoustic signal as cepstral coefficients because we believe that this representation is a good model of the output of the human auditory apparatus. This is an informed, but ultimately subjective decision that might not be appropriate for other scientific investigations, e.g., the processing of sounds by birds or bats. This point is discussed in more detail in Sect. 9.

In what follows we first discuss related work in Sect. 2 and describe our experimental setup in Sect. 3. We then demonstrate the difficulty of segmentation by showing that a whole-sequence method that computes similarity between sequences works reasonably well for artificial data, but does not succeed in learning to segment natural speech (Sect. 4). In Sect. 5, we define several baseline segmentation algorithms that operate on the level of a single frame. In Sect. 6, we present the *extended context model* that succeeds in learning a rudimentary form of segmentation by incrementally extending the context that is used to categorize an interval of speech. Baseline methods and the extended context model are evaluated in Sect. 7. The segmentation computed by the extended context model is the basis for a successful correspondence simulation in Sect. 8. Section 9 analyzes and discusses the results and Sect. 10 presents our conclusions.

2 Related Work

2.1 Approaches to Segmentation Based on Representations of the Acoustic Signal as a Symbolic Sequence

Most work on segmentation of speech in cognitive science has assumed a coarse-grained representation of the input signal as a sequence of either syllables, phones or letters (Brent 1999). We call this approach the *symbolic sequence* approach. In this approach, the unsegmented representation of "five black dogs" might look like [faɪvblækdɔgz] or `fiveblackdogs`.

A variant of symbolic sequences represents speech as a sequence of *phonetic feature vectors* using features such as *voiced* and *nasal*. For our purposes, this type of feature representation is equivalent to symbolic sequences because the features are usually derived from phones (not from the speech signal) and there is a simple correspondence between phones and feature vectors. For example, a bilabial nasal corresponds to [m].

Segmentation methods in this line of work address the problem of segmenting the input sequence into *syllables* or *words*, e.g., [faɪv.blæk.dɔgz] or `five.black.dogs`. When the problem is couched this way, the difficulty of the problem lies in finding the segment boundaries; the classification of the segments directly follows from the segmentation boundaries. The assumption is that all instances of the five letter sequence `black` belong to the same class of syllables or words; and that all other sequences of letters do not belong to this class. The challenge of learning the phones/letters

themselves is not addressed in this approach. This is the challenge we focus on in this paper.

Following Brent (1999), we distinguish the following types of strategies in symbolic sequence segmentation: the *utterance-boundary strategy*, the *predictability strategy*, the *word-recognition strategy*, and the *boundary-cue strategy*.

Utterance-boundary strategy models hypothesize boundaries based on sequences of lower-level segments (e.g., phones) that are characteristic for the endings of the segments to be identified. For example, Xanthos (2003) presents an incremental method for speech segmentation that finds word boundaries, using a symbolic sequence approach based on sub-word segments (phonemes).[1]

We believe that the utterance-boundary strategy is not appropriate for phone acquisition in the context of correspondence learning because there is little information about utterance boundaries available in the initial stages of phone acquisition. On the level of frames (2–10 milliseconds long), the number of utterance-final events is extremely small, making it difficult to generalize on the acoustic level. [2]

The *predictability strategy* is based on predicting boundaries at places in the signal where prediction of the next segment based on the current context is hard. The assumption is that the possible sequences of segments within words are more restricted than the possible sequences across word or syllable boundaries. In our experiments below we will evaluate two versions of the predictability strategy, transition probability and entropy.

The *word-recognition strategy* attempts to learn an inventory of words and posits a segmentation boundary wherever one word ends and another one begins. Some models implementing this strategy are based on the assumption that all, or at least a sufficient number of the units to be recognized (i.e., the words) are first heard in isolation (Christophe et al. 1994); and that remaining segments of the speech stream between recognized words are also words. A segmentation method for continuous speech is presented by de Marcken (1996) within the broader context of unsupervised language acquisition. The algorithm learns a lexicon from unsegmented input which is then used for segmentation. The learning procedure operates on a symbolic sequence of transcribed speech, which is automatically obtained from audio input by a separate system, external to the actual learning module. More recently, unsupervised methods, in particular, Bayesian methods like Dirichlet processes (Goldwater et al. 2009), have been investigated that learn to segment by way of learning an inventory of words. Frank et al. (2007) provide a review and a comparison of some related segmentation models with experimental results from an artificial language learning task with adults.

The main problem of the word-recognition strategy is that its basic assumption, namely that there is a stable higher-level unit like the word, is not correct for phone acquisition. For example, in our corpus (Schweitzer et al. 2004) fewer than 20% of the

[1] Related work by Xanthos (Goldsmith and Xanthos 2009) explores a range of methods for establishing whether it is possible to automatically infer whether segments in a data sample are vowels or consonants (in addition to examining vowel harmony and phonotactic induction). Segmentation, however, is not the focus. Similarly for related work by Vallabha et al. (2007), who propose an unsupervised algorithm "for learning the categories from a sequence of vowel tokens" from infant-directed speech.

[2] Preliminary experiments with an utterance boundary strategy confirmed this. The utterance boundary strategy performed worse than all the methods evaluated in Table 3.

instances of [t] have sequences of mel-frequency cepstral coefficient (MFCC) frames that are identical to another instance of [t] (see Sect. 3). Thus, while all instances of a word in a letter representation are identical, this is not true for the instances of a labeled phone in a representation derived from speech. A related problem is that the number of possible words in language is not bounded. In contrast, there is a comparatively small inventory of phones although different granularities of the inventory might be appropriate in different contexts.[3] Work in the Bayesian framework assumes an infinite vocabulary (Goldwater et al. 2009). For these reasons, we will not consider word recognition as a possible approach to phone acquisition.

The *boundary cue strategy* attempts to segment the speech stream based on reliable acoustic cues found, for instance, in stressed syllables or at phonetic junctures, which infants are hypothesized to have a disposition to search for and locate in continuous speech (Cairns et al. 1997; Christophe et al. 1994; Jusczyk 1999). For example, Christophe et al. (1994, p. 1570) write:

> [...] as early as 1939, Trubetzkoy described a number of potential cues that could demarcate words: He mentioned allophonic variation [...], lexical stress [...], vowel harmony, tone phenomena, among others. [...] allophonic cues were shown to influence English subjects' parsing of pairs of words such as "gray tie" versus "great eye" [...] The generic hypothesis is that something perceivable signals word boundaries [...]

Many authors agree that there must be such cues and that a rough partitioning of speech based on these cues is performed by infants during language acquisition. However, there is no agreement what exactly these cues are (although the ones listed by Trubetzkoy are all plausible candidates) and how they would be used computationally (cf. Jusczyk 1999).

In a sense, our contribution in this paper can be viewed as defining a segmentation algorithm based on boundary cues and thereby confirming the underlying hypothesis of these models: learning a rough segmentation of the speech signal into a sequence is possible.

2.2 Approaches to Segmentation Based on Direct Representations of the Acoustic Signal

2.2.1 Semi-Supervised Approaches

A number of approaches to the speech segmentation problem have been investigated that use a representation that is directly derived from the speech signal; most of this work has been done in the fields of automatic speech recognition (ASR) and concatenative speech synthesis. Phone acquisition, however, has received little attention in this area, where the standard approach is to train recognizers on a pairing of an audio

[3] For example, one may want to distinguish different subcategories for the phone [k] (syllable-initial vs. syllable-final, before front vs. back vowel etc), but different instances of [k] in syllable-initial position before front vowels should be assigned to the same phone. See the notes on terminology in the introduction.

signal and a transcription (Huang et al. 2001). Thus, the term *speech segmentation* in ASR and speech synthesis often refers to the process of aligning a given transcription with the speech signal; this is a semi-supervised setup in the sense defined above: human labels are available, but not for the task to be solved. In these approaches, the transcription into labeled phones, syllables or words assumes a prior definition of these categories—even if the "semi-supervision" is only used to initialize a model that is then refined in an unsupervised fashion (cf. Ljolje et al. 1997; Toledano et al. 2003; Varadarajan et al. 2008).

Other work on speech segmentation in ASR does not evaluate the quality of segments learned in an unsupervised fashion even though their benefit in an application is shown (Stouten et al. 2008; van Segbroeck and Van hamme 2009).

A number of approaches to language acquisition in the phonetic realm frame the problem as recognizing *entire words* in longer utterances before *phones* are acquired (ten Bosch et al. 2008; Roy and Pentland 2002; Werker and Curtin 2005). Presumably, phones would then be derived in a subsequent step from the recognized words—although some authors raise the possibility that phones and segmentation into phones do not play any important role in language acquisition and that all learning happens on the level of words and utterances. We discuss these approaches in this section on semi-supervised work because no segmentation labels are provided to the learner (thus, the approach is not supervised), but instead semantic labels are available that identify different utterances as referring to the same object. Here, we focus on ACORNS (ten Bosch et al. 2008), a project that has done the most significant work so far on realistic computational modeling of child language acquisition taking the "whole-utterance" approach.

Methods developed in ACORNS can recognize words in longer utterances with high accuracy (ten Bosch et al. 2008). The feature representation of utterances is constructed in an unsupervised fashion, followed by supervised training of a word recognizer on utterance-label pairs. In a number of different experiments the speech data consisted of between 10 and 20 distinct words, e.g. 11 words in (Altosaar et al. 2010) and 13 words in (ten Bosch et al. 2008). While learning of phones may not be necessary if the inventory of words is as small as 20, it is unclear whether the methods developed in ACORNS can be used for larger vocabulary sizes or in environments where a non-noisy mapping of utterances to semantic labels is not available.

Both our approach and ACORNS start with a representation directly derived from speech and at some point in the learning process use utterance-label pairs. The main conceptual difference is that we learn a segmentation first, in an unsupervised fashion, and then represent utterances as sequences of segments. In contrast, ACORNS uses an HAC (histogram of acoustic cooccurrences) representation: utterances are first represented as vectors of vector quantization label cooccurrences and then reduced in dimensionality (Driesen et al. 2009). In this additive representation, for a given cooccurrence of two labels, the information whether it occurred at the beginning or end of the word is not retained. Thus, the representation used in ACORNS is non-segmented and non-sequential. The ability of the ACORNS model to recognize new instances is then interpreted as evidence that segmentation is perhaps not needed. It remains to be shown that this approach will scale from fewer than 20 words to a more realistic vocabulary size. We believe that the relatively impoverished representation of

utterances used by ACORNS, a representation that does not retain temporal information at multiple scales, would require more utterance-label training data than is available in the child's input during language acquisition. The training sets used in ACORNS (e.g., 3000 utterances in (Driesen et al. 2009)) are larger by about a magnitude than what is used in the approach to correspondence learning described in Sect. 8. Consequently, even though ten Bosch et al. (2008) use a somewhat richer experimental set-up than we do (e.g., a multi-speaker corpus and multiple-word utterances, as opposed to our single-speaker single-word corpus) we believe that their underlying assumption might not be correct and that our segmentation approach should scale well to larger vocabularies.

Also relevant in the context of semi-supervised learning are proposed learning systems which incorporate articulatory or visual information (e.g. Blackburn and Young 1996; Roy and Pentland 2002; Coen 2006). However, these systems are not primarily concerned with segmentation and correspondence learning.

2.2.2 Unsupervised Approaches

The methods discussed so far are supervised or semi-supervised. There is a fundamental problem with such models. In the words of Cairns et al. (1997, p. 144): "explicit training is not normally a part of human development, so we need to explain how the model can come to be part of the human language processor"; we would include approaches here that predefine the inventory of phones, even if it is not in the context of explicit supervised learning. It is true that models of adult speech segmentation should incorporate top-down mechanisms of information flow, e.g., by generating expectations based on the previous context or the lexicon (cf. McQueen 1998; Wade et al. 2010); but the initial unsupervised structuring of the acoustic signal in a phone acquisition model has to rely on bottom-up mechanisms without recourse to extensive knowledge about the language. Most relevant to segmentation and phone acquisition are therefore approaches that are unsupervised. To date, however, these approaches have had only limited success and in many cases the evaluation has not been, in our opinion, sufficiently rigorous. We now review a number of studies in this vein.

Lin (2004, 2005) describes an approach for learning phones directly from the speech signal, in which phonetic features are automatically learned, but with mixed results; e.g., half of all [b]'s are labeled as nasal. Segmentation is not addressed. Sharma and Mammone (1996) present an unsupervised segmentation algorithm that tries to find the "optimal number of sub-word segments". However, they test it on only a single word. Gold and Scassellati (2006) compute MFCC representations for a set of thirty utterances. They derive acoustic features from the speech signal, but do not detect phone boundaries. Scharenborg et al. (2007) demonstrate the difficulty of attempting to learn segmentation in an unsupervised fashion based on MFCC vectors. They focus on the challenges faced by the infant in learning segmentation as opposed to an evaluation that sheds light on the difficult precision-recall trade-off inherent in segmentation - positing few boundaries results in many false negatives, positing many boundaries results in many false positives. In a 2010 follow-up study, they conclude that segment duration, inherent segment dynamics or the adjacency of similar segments are fundamental problems which unsupervised segmentation algorithms have

to solve. To address these, they propose the incorporation of "automatically derived top-down information"—e.g. automatically derived broad classes based on clustering—in addition to the usual unsupervised "bottom-up" segmentation (Scharenborg et al. 2010). This is in line with our approach using clustering (see Sect. 3). Miller et al. (2009) carry out segmentation experiments based on automatically tokenized speech spectrogram data. They do not evaluate how well their method recognizes boundaries between phones.

While these cases have provided useful insights, evaluation has tended to be limited. Our evaluation, however, is more comprehensive as it compares our segmentation approach to multiple baseline models and evaluates it in the context of the actual problem that needs to be solved by the child, the correspondence problem.

2.3 Correspondence and Segmentation in Language Acquisition

There are numerous studies that investigate at what age infants have established correspondence, but there are to our knowledge no computational models that show that a specific algorithm, working on real speech data, can successfully establish correspondence for phones. Furthermore, there is no evidence that correspondence, e.g., the mapping between articulatory muscle movements and particular sounds, is innate.

Many authors have addressed the problem of how representations of linguistic units like phones and words emerge and how they are extracted from the continuous speech stream—see, for example, the empirical studies by Saffran et al. (1996) and Christophe et al. (1994). Jusczyk (1999) reviews the phonetics and cognitive science literature on the topic and identifies a number of issues that are not addressed by currently existing limited models of word extraction.

It is important to point out that (i) mature adult speech segmentation and (ii) the initial acquisition of a basic segmentation capability by the child may be very different. For example, mature speakers will obviously make use of top-down constraints to guide segmentation: knowledge of words, grammaticality, and plausibility. Infants cannot do this in the first stages of language acquisition. Thus, there may be little commonality between the mechanisms used in adult speech segmentation and the mechanisms for learning segmentation based only on information sources available to the infant (Cairns et al. 1997).

Our main motivation for modeling segmentation computationally is its hypothesized importance for correspondence. With regard to children's ability to segment speech, there is considerable evidence that children learn to do this at an early age. At 4 months of age, infants already possess some knowledge about the correspondence between auditory and visual speech information (Kuhl 1988; Kuhl and Rivera-Gaxiola 2008), making it likely that a rudimentary form of segmentation has been learned. Around the age of 7 months, infants begin to segment words from the continuous speech stream (Jusczyk 1999). By the age of 9 months, infants seem to have developed a sensitivity to the phonotactics and lexical stress patterns of their native language (Mattys et al. 1999). Usually, this sensitivity is assumed to rely on atomic units of speech perception, which again would imply that these atomic units have already been learned by this age.

Research has shown that there is a lag between the early development of abilities in speech perception and the later development of the corresponding articulatory abilities (cf. Kuhl 1987). This indicates that articulation is perhaps a contributor to segmentation, but that segmentation should not be entirely driven by articulation. Arguably, articulation does play an important role in learning segmentation during later phases of language acquisition when articulatory gestures are helpful in determining boundaries between phones (the link I/AC-I/AR in Fig. 1). However, exploration of the use of articulatory information is beyond the scope of this article, and in the experiments that follow we focus purely on acoustic information.

3 Experimental Setup

Formally, we pose the *speech segmentation problem* for correspondence learning as follows:

- Let the probe $x = [t_s, t_e]$ be a (short) interval of speech that is to be segmented.
- Let y be a (large) memory that stores speech experienced in the past.
- A segmentation of the probe x based on y is a sequence of segmentation boundaries $t_1 < t_2 < \cdots < t_n, n \geq 1, t_s < t_1, t_n < t_e$ that segments x into $n + 1$ segments that can be successfully used in correspondence learning as defined in Sect. 1. The segmentation boundaries are selected based on a computational analysis of the distribution of acoustic sounds in y.

This definition of the problem is fundamentally different from most segmentation problems in psycholinguistic computational modeling in that there is no obvious base alphabet in which the input can be represented. The memory y also is fundamentally different from memory-based approaches like Exemplar Theory (e.g. Johnson 1997; Goldinger 1997, 1998; Pierrehumbert 2001) or k-nearest neighbor classification: It is not a memory of items, but instead one long sequence (cf. Wade et al. 2010). Representing memory as a set of segmented items would assume a solution to the very problem we are trying to solve.

In the experiments described below, we use three types of corpora: Constant-shift corpora, variable-shift corpora and a corpus of German speech.

3.1 Constant-Shift Corpora

These corpora were constructed to make the segmentation task as easy as possible while providing some insight into properties of different segmentation methods. We first randomly generated a 12 dimensional target segment that has length l_p, the *target length*. We also chose a second parameter l_Δ, the *shift*, that specifies the distance between two successive occurrences of the target segment. A constant-shift corpus for the parameters l_p and l_Δ was then generated by copying the target segment, generating $l_\Delta - l_p$ random frames (or vectors, intended to model MFCC vectors) and repeating this process up to 1,000,000 frames, the desired length of the corpus. The vectors were then length-normalized.

Each constant-shift corpus thus generated was then a memory y for a particular experiment. The probe x for that experiment was the target segment padded with additional random frames to a maximum length of 100. Since there is ample evidence in y that at the end of the target segment possible continuations diverge, it should be easy to segment x into the target segment and its second randomly generated part.

3.2 Variable-Shift Corpora

Variable-shift corpora were generated in the same way as constant-shift corpora except that l_Δ was not constant, but a random variable whose value was randomly selected from the interval $[l_p + 1, \alpha \cdot l_p]$. This interval was chosen to prevent target overlap, immediate repetition and unrealistically long segments. We experimented with various parameter combinations, and report results for a maximum shift factor $\alpha = 17$ in the following section. Variable-shift corpora are slightly more realistic than constant-shift corpora since target segments in real speech do not occur with intervening intervals of fixed length. But the task of finding the target segment in the probe is still easy compared to speech.

3.3 IMS Unit Selection Corpus

We also use the IMS unit selection corpus (Schweitzer et al. 2004), a corpus of German speech, recorded by a professional male and a professional female speaker, and sampled at 16,000 Hz. 13-dimensional MFCCs were computed for the 2776 original speech files of the male speaker-part by means of the MATLAB Auditory Toolbox (Slaney 1998). MFCCs are considered to be a suitable approximation of human speech perception (Morgan et al. 2004) and have been successfully applied in a number of experiments involving auditory models and speech representations as well as in technological applications like ASR. We set the parameter `frameRate` of the `mfcc` function of the Auditory Toolbox to 500, corresponding to a 2 ms window shift. The remaining parameters were not changed from their default values. Only the last 12 components of the MFCC vectors were used. The vectors were length-normalized. Silence was excluded from the experiments except for a 20 ms interval before and after each speech interval. This is based on the assumption that a separate module is available to the learner that distinguishes speech from non-speech.

We computed a vector quantization (Gersho and Gray 1991) of the MFCC vectors. The vector quantization was performed using bisecting k-means (Steinbach et al. 2000) and Euclidean distance. 1,000 clusters were generated and reflect dense regions of MFCC space.

The evaluation is based on the phonetic transcription of the IMS corpus, which consists of a total of 107,209 labeled phones that are between 1 and 276 frames long. These labels are only used for evaluation in phone acquisition, not for learning. The labels are also used in correspondence learning as described in Sect. 8.

As a result, the IMS corpus is represented as a sequence of 4,310,124 frames, each consisting of an MFCC vector, a cluster identifier and a linguistic label. Except for the whole-sequence experiments in Sect. 4, the corpus is split into a training set (frames

1–3,000,000) and a test set (frames 3,000,001–4,310,124). All unsupervised learning is performed on the training set. For example, the probabilities needed in some of the segmentation methods described below are estimated on the training set. In the evaluation (Sect. 7), the 1.3 million frames of the test set are treated as new unseen data that are segmented using the parameters learned from the training set.

3.4 Evaluation Measures for Segmentation

Traditionally, segmentation has been evaluated with respect to a gold standard of correct segmentation boundaries, e.g., by measuring precision and recall for the task of finding the correct boundaries (cf. Aversano et al. 2001; Qiao et al. 2008; Goldwater et al. 2009; Scharenborg et al. 2010). Such traditional segmentation measures cannot be used for evaluating phone acquisition because they either do not assign segments to phones (each segment being equally unrelated to all other segments); or because they classify segments in a trivial way: two segments are assumed to be in the same class if their sequences are identical and vice versa. Phone acquisition is an integrated segmentation and classification task: both segment boundaries and categories of the segments have to be learned. As mentioned above, we use an MFCC representation which is directly derived from an audio recording of speech. Real speech data are highly variable. For this reason, almost every segment of 20 frames or more in the speech stream is unique. Thus, the simple criterion of identity that is used in the symbolic sequence approach is not applicable when trying to group segments into classes in phone acquisition.

To address this problem, we *cluster* segments and interpret the clusters as induced phones where the number k of clusters is a parameter of the evaluation. This cluster-based evaluation is a more direct evaluation of what knowledge about phones means: it involves being able to determine where one phone ends and the next one begins; but of equal importance is the ability to determine whether two segments belong to the same phone or not.

The clustering of segments was performed using bisecting k-means (Steinbach et al. 2000). In two separate evaluation runs, two different representations of segments are used. First, segments were represented as normalized 1000-dimensional vectors, where dimension i represents the count of frames of MFCC cluster i in the segment. For example, if a segment consists of 20 frames, with the first 5 frames assigned to MFCC cluster 16 and the last 15 frames assigned to MFCC cluster 226, then the vector $\vec{v}$ of the segment has $v_{16} = .25$, $v_{226} = .75$ and a value of 0 for the other 998 dimensions. This evaluation is a true evaluation of the joint segmentation and classification task that better reflects the phone acquisition task faced by children as no gold standard information is available.

We also compute evaluation numbers for a second clustering in which each segment is represented as a vector of gold standard labels. For example, if a segment consists of 20 frames, with the first 5 frames assigned to [ʃ] and the last 15 frames assigned to [æ], then the vector $\vec{v}$ of the segment has $v_{[s]} = .25$, $v_{[æ]} = .75$ and a value of 0 for the other 61 dimensions. (The total size of the inventory of labeled phones in the corpus is 63.) This evaluation sheds light on how well segmentation boundaries are

learned, but is not a good model of actual learning by the child since the unrealistic assumption is made that each frame has been classified correctly as belonging to one of 63 linguistic phones.

As with the vector quantization described in Sect. 3, the evaluation clustering is performed on the training set. Test set segments are assigned to the clusters found in a subsequent step.

The labeling of the IMS corpus is not ideal for evaluating our task because many instances of labeled phones consist of several distinct segments when looking at acoustic measures. For example, an interval that is labeled as being an instance of the stop [t] may correspond to three subintervals: silence, plosion and frication. Similarly, the German diphthong [aɪ] often consists of three subintervals, resembling [a], [ɪ] and a transition between the two sounds in between. Conversely, intervals bearing the same phonetic label (e.g., [k] in [ki] vs. [ku]) can be acoustically quite different.

Thus, given a language and a signal, there is no one true inventory of phones of the language and no one true labeling of the signal based on these phones. Instead, there are different labelings of different granularity. A fine-grained labeling may distinguish the three phases of [t] and [aɪ]. In a coarse-grained labeling, these two sounds may each be segmented and classified as single phones. Ultimately, the granularity of the labeling should be decided based on the application of the labeling, the learning of correspondence in this paper.

The granularity of the labeling corresponds to the notion of *boundariness*—we view boundaries between phones as graded. A weak boundary (corresponding to a small value of boundariness) will only be chosen as a boundary for a fine-grained labeling. A strong boundary (corresponding to a high value of boundariness, e.g., a plosion) will be chosen by both fine-grained and coarse-grained labelings.

We incorporate these considerations into our evaluation by introducing a parameter a whose value governs the degree of granularity. The value of a is roughly the average size of a segment. More precisely, we compute the segmentation of the corpus C based on a by choosing the largest segmentation threshold θ that creates more than $|C|/a$ segments where $|C|$ is the number of frames in C. To ensure a fair comparison between segmentation methods, ties between boundaries are broken randomly to get exactly $\lceil |C|/a \rceil$ boundaries. This procedure guarantees that the average length of a segment is approximately a. The segmentation function σ determines for point t whether a segmentation boundary occurs before t ($\sigma(t) = 1$) or not ($\sigma(t) = 0$). For the segmentations we investigate in this paper, we always define σ in terms of b and θ as follows:

$$\sigma_{b,\theta}(t) = 1 \text{ iff } b(t) \geq \theta$$

where b is a boundariness function that computes for each t the boundariness just before t. As a final step, we eliminate overly short or long segments. As will be discussed in Sect. 7, average segment lengths a are in the interval [30 frames, 100 frames] or [60 ms, 200 ms]. We eliminate segments that have less than half the length of the lower bound of the interval (i.e., segments with length $l < 15$) and segments that have more than twice the length of the upper bound of the interval (i.e., segments with

length $l > 200$). We view this as an implementation of our prior knowledge of what the desired range of durations of phones is.

We use two clustering measures to evaluate the quality of the induced phones: purity (P) and Adjusted Rand Index (ARI).

To compute purity (Strehl 2002), each induced phone ω_l is assigned to the labeled phone c_i whose frames are most frequent in ω_l, and then the accuracy of this assignment is measured by counting the number of correctly assigned frames and dividing by the total number of frames N:

$$\text{purity}\ (\Omega, C) = \frac{1}{N} \sum_{l} \max_{i} |\omega_l \cap c_i| \tag{1}$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of induced phones and $C = \{c_1, c_2, \ldots, c_J\}$ is the set of labeled phones. We interpret ω_k as the set of frames assigned to ω_k and c_j as the set of frames labeled with label c_j in Eq. (1).

Bad clusterings have purity values close to 0, and a perfect clustering has a purity of 1. Purity is an intuitive measure of cluster quality, but high purity is easy to achieve when the number of clusters is large—in particular, purity is 1 if each frame gets its own cluster. For this reason, we also use ARI because it assesses the quality of the trade-off between the purity of the clustering and the number of clusters.

Adjusted rand index (Hubert and Arabie 1985) views clustering as a series of decisions, one for each of the N^2 ordered pairs of frames in the corpus. Let i_f (i_g) be the induced phone of frame f (g) and l_f (l_g) its labeled phone. A pair of hypothesis frames (f, g) is considered a true positive (TP) if $i_f = i_g \wedge l_f = l_g$, a true negative (TN) if $i_f \neq i_g \wedge l_f \neq l_g$, a false positive (FP) if $i_f = i_g \wedge l_f \neq l_g$, and a false negative (FN) if $i_f \neq i_g \wedge l_f = l_g$. Adjusted rand index is then defined as follows:

$$\text{ARI} = \frac{2(\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN})}{(\text{TP} + \text{FP})(\text{FP} + \text{TN}) + (\text{TP} + \text{FN})(\text{FN} + \text{TN})}$$

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives and false negatives, respectively. Adjusted rand index is zero for a random clustering and 1.0 (the maximum value) for a perfect clustering.

The boundariness functions in this paper are deterministic, but the clustering of segments into induced phones is not. We therefore report averages of 10 trials and adopt a hypothesis testing framework.

4 Whole-Sequence Model

In this section, we show that a simple baseline model that directly compares entire sequences to each other—without any analysis of individual frames—does not learn speech segmentation in an unsupervised fashion. We first show that this model has some success on artificial, non-noisy data. We then show that it does not work on more noisy natural speech.

4.1 Whole-Sequence Segmentation Criteria

We went through a number of iterations in devising an unsupervised segmentation algorithm. Our initial intuition was that a segmentation point t_i in $x = [t_s, t_e]$ is a point with the following two properties:

- There are many sequences in y with high similarity to $[t_s, t_i]$.
- Extending the interval by ϵ to $[t_s, t_i + \epsilon]$ does not increase the overall similarity of the sequence with the memory.

To formalize this segmentation method, we first define correlation and cumulation. As described in Sect. 3, probe x and memory y are represented as sequences of MFCC vectors.

The raw correlation score r between a prefix of x and a subsequence of y is defined as follows:

$$r(x, y, m, n) = \sum_{l=1}^{12} \sum_{i=m}^{n} x_{l,i-m+1} y_{l,i} \tag{2}$$

where x_{lk} (resp. y_{lj}) is the l^{th} component of the MFCC vector with index k in x (resp. j in y). $r(x, y, m, n)$ is a measure of how well the first $n - m + 1$ frames of x match the subpart of y from index m to index n. m, n, x must satisfy $n - m + 1 \leq |x|$.

The normalized correlation score ρ between a prefix of x and a subsequence of y is defined as follows.

$$\rho(x, y, m, n) = \sum_{l=1}^{12} \frac{\sum_{i=m}^{n} x_{l,i-m+1} y_{l,i}}{\sqrt{\sum_{i=m}^{n} x_{l,i-m+1}^2} \sqrt{\sum_{i=m}^{n} y_{l,i}^2}} \tag{3}$$

The cumulation function computes a score that is accumulated along the length of x, either for $f = r$ or for $f = \rho$:

$$C(x, y, f, k) = \sum_{i=1}^{|y|-|x|+1} f(x, y, i, i + k - 1) \tag{4}$$

The derivative of the cumulation function indicates how much additional similarity can be found when extending the prefix of x gradually, step by step:

$$C'(x, y, f, k) = C(x, y, f, k) - C(x, y, f, k - 1) \qquad (k \geq 1) \tag{5}$$

As an extreme example, suppose that x is [bali:li:] and that y is a concatenation of 1,000 [ba] syllables. Let k be the length in frames of the first two phones in [bali:li:]. Then $C'(x, y, r, k)$ will be high because as we extend the prefix of x from the beginning of the first [b] to the end of the first [a], each of the 1,000 instances of [ba] in y will contribute more and more similarity to C. Thus, the value of the derivative will be high. However, $C'(x, y, r, k + 1)$ will be close to zero or negative. Once we

have passed the end of the first [a], none of the 1,000 instances of [ba] in y (and no other subsequences in y) will contribute additional similarity to C. In fact, negative correlation may decrease overall similarity. Thus, the value of the derivative will be small or negative.

Our initial hypothesis was that minima of C' could serve as segmentation boundaries. We quickly found in experiments on natural speech data that this is not the case. The reason is that when working with real speech data, the overwhelming majority of subsequences of y have almost no or negative similarity with the prefix of x. In the above example, [ba] would be compared with all diphones in y, but most of these (e.g., [aʊ], [at], [lt]) are not relevant for making segmentation decisions within x.

To ensure that a segmentation decision is based only on sequences with a minimum similarity, we introduce a threshold value θ. Any sequence in y with similarity below θ is not considered in segmentation decisions. To formally define this modified method, we first define mean and standard deviation of a set of values. We use the symbols μ and σ for their estimates:

$$\mu(S) = \frac{\sum_{v \in S} v}{|S|} \tag{6}$$

$$\sigma(S) = \sqrt{\frac{\sum_{v \in S} (v - \mu(S))^2}{|S| - 1}} \tag{7}$$

The set A of all "above-threshold" values is then defined as a subset of the set V of all values as follows:

$$V(x, y, f, k) = \{f(x, y, i, i + k - 1) | 1 \leq i \leq |y| - |x| + 1\} \tag{8}$$

$$A(x, y, f, k) = \{v \in V(x, y, f, k) | v \geq \mu(V(x, y, f, k)) + \lambda\sigma(V(x, y, f, k))\} \tag{9}$$

That is, $A(x, y, f, k)$ is a set of values, each of which corresponds to the similarity (computed by f) between a prefix of x and a highly similar subsequence of y. We set $\lambda = 1$.

We can now define the boundariness function m' we use:

$$m(x, y, f, k) = \mu(A(x, y, f, k)) \tag{10}$$

$$m'(x, y, f, k) = m(x, y, f, k) - m(x, y, f, k - 1) \quad (k \geq 1) \tag{11}$$

Negative values of the derivative m' indicate that the average similarity of the most similar subsequences in y is dropping off. We would expect this to be the case at segmentation boundaries because up to a segmentation boundary the similarity with a large number of highly similar subsequences of y is growing; after the segmentation boundary these subsequences are continued in different ways (e.g., [ba] might be followed by [ga], [tiː] or [luː]) and therefore the growth in average similarity will drop or become negative.

In addition, we also evaluated $\sigma(A(x, y, f, k))$ as a segmentation function. The motivation was that a prefix of x that occurs many times in y will have consistent similarity scores in a small range with its most similar nearest neighbors in y.

This translates into small standard deviation. In contrast, at a point within x that can be continued in a number of different ways (at the end of the [a] in [ba], we can continue the sequence with [ga], [ti:], [lu:] etc.), high variability of the scores of the highest similarity subsequences in y will occur.

4.2 Experimental Results

Results for constant-shift corpora are shown in Fig. 2.

We show a number of representative values for target length l_p and shift l_Δ. Although the results illustrated in Fig. 2 on artificial data are positive, the results of this section on real speech are essentially negative (see Fig. 5), therefore we do not perform an extensive evaluation for all possible values. The graphs are based on computations using a mixed MATLAB/C implementation.

The horizontal axes show the length of the probe prefix in frames. Recall that the length of a frame is 2 ms. The dotted vertical lines are true segmentation boundaries, aligned to the beginning of the target segment in the corpus and the probe. Each X in the upper part of the graphs marks a target segment in the corpus y. The label "…" refers to non-target material, i.e., the random parts of the sequence. For the target segment, its true length l_p is marked by the first dotted vertical line in each graph, and the remaining frames up to the maximum probe length 100 correspond to the padded random frames in the probe x.

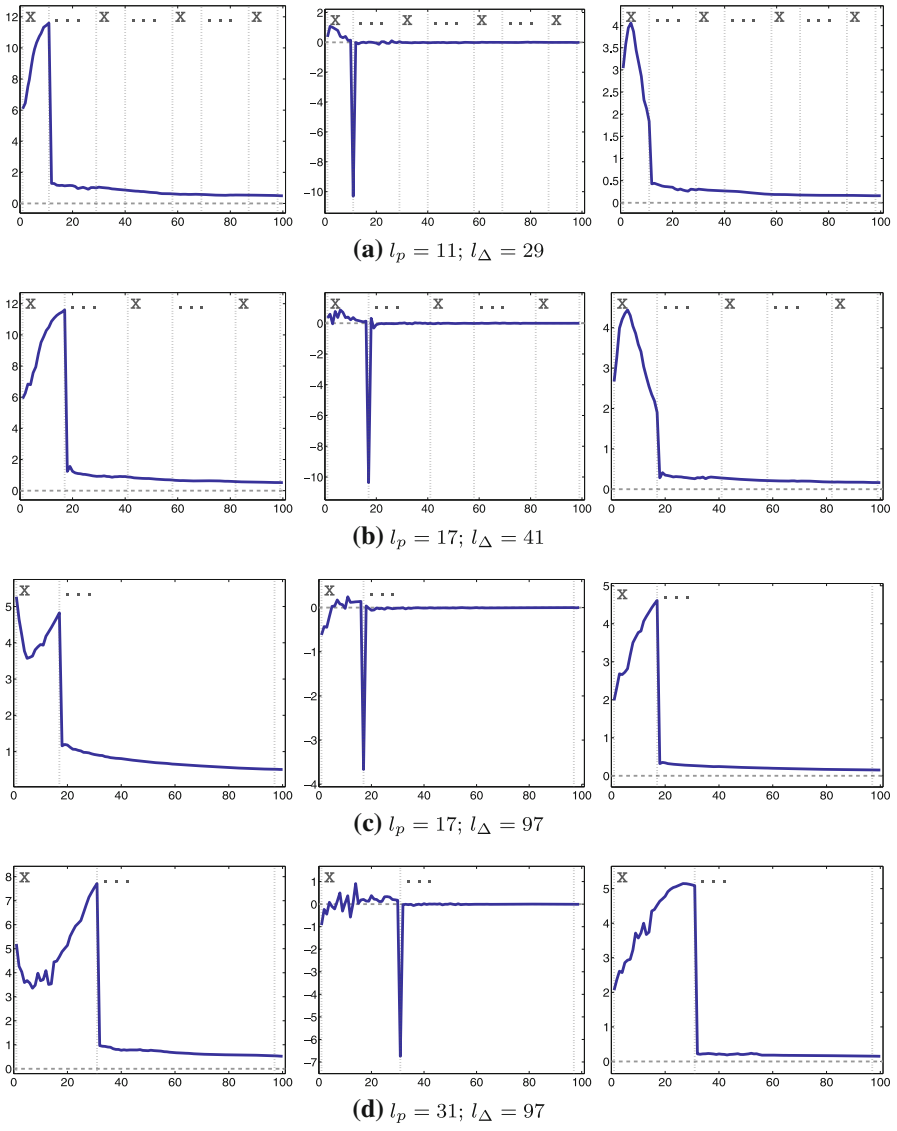
Minima of the derivative $m'(x, y, f, k)$ correctly indicate segmentation boundaries. A steep drop of $m(x, y, f, k)$ and $\sigma(A(x, y, f, k))$ from a maximum correctly predicts the segmentation boundaries. This is a sample representative of a larger number of experiments we ran: minima of m' and a step-down of σ are reliable indicators of segmentation boundaries.

Results for variable-shift corpora are shown in Fig. 3. Minima of m' again indicate segmentation boundaries reliably, as do the maxima of m and σ preceding the step-down. However, all the values have a somewhat lower magnitude in comparison to the results of the constant-shift corpora. Considering these results, all three functions could be seen as possible candidates for speech segmentation.

Figure 4 shows the results of two experiments for a nontarget probe, i.e., a probe that did not contain a prefix that corresponds to the target segment. With such a random probe, the results always look similar to the ones shown here, independent of the actual parameter settings: There seems to be no correlation between the function values and the segmentation boundaries. This means that false positive boundaries can be avoided for the constant-shift and variable-shift cases.

Figure 5 shows results for speech data. We computed m, m' and σ for a number of probes for $\lambda = 2$ and $f = \rho$ (no parameter combination we tried produced better results). Probes were chosen to begin at a phone boundary following a stretch of silence (i.e., at word- or utterance-initial phones). For example, the top graphs in the figure are for the first 100 frames of a sequence of two phones, [li:].

Results for m, m' and σ are disappointing. Even though they have some promise for segmenting non-noisy artificially generated data, they do not seem to be able to find segmentation boundaries reliably for speech.



(a) $l_p = 11;\ l_\Delta = 29$

(b) $l_p = 17;\ l_\Delta = 41$

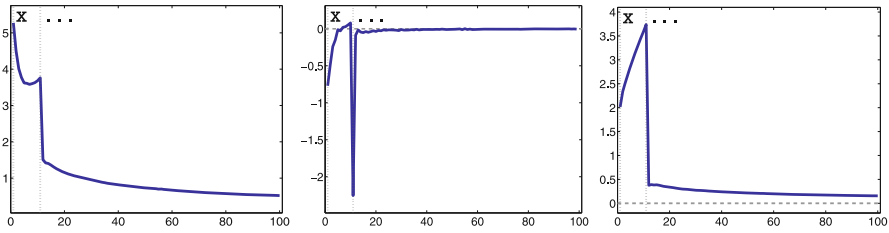
(c) $l_p = 17;\ l_\Delta = 97$

(d) $l_p = 31;\ l_\Delta = 97$

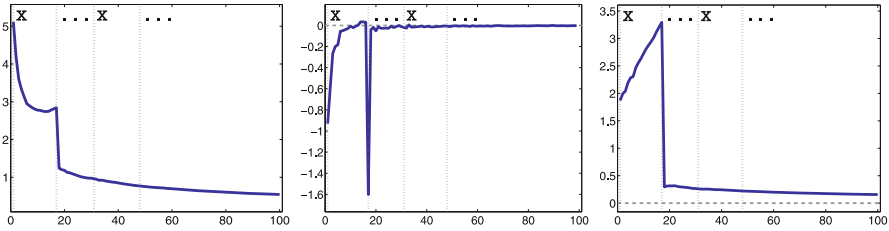
Fig. 2 Results for constant-shift corpora. x-axis: length of probe prefix in frames. y-axis: $m(x, y, \rho, k)$ (the four left panels), $m'(x, y, \rho, k)$ (the four middle panels), and $\sigma(A(x, y, \rho, k))$ (the four right panels). *Dotted vertical lines* are true segmentation boundaries

4.3 Summary

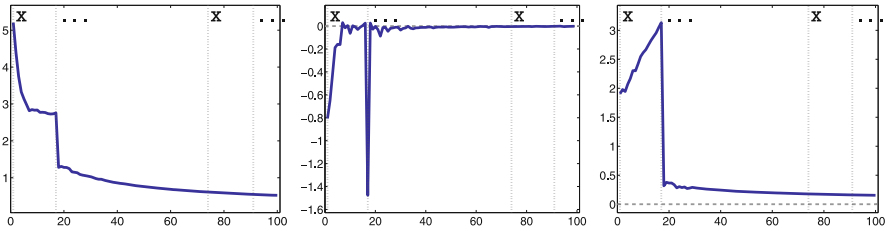
We have had some success in developing a segmentation criterion for non-noisy artificial data. However, this criterion did not perform well for speech. The memory y of the artificial data was constructed to contain the probe many times. It is perhaps not surprising that for speech data, where the probe also occurs many times in the memory,



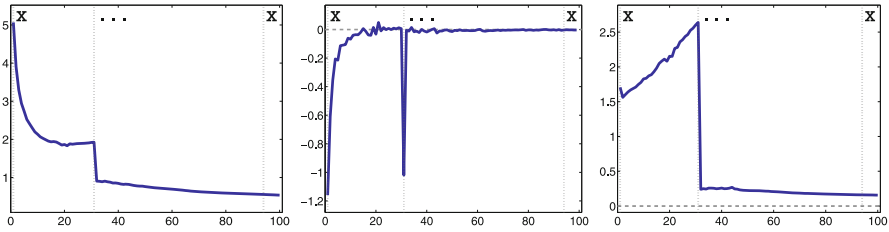
(a) $l_p = 11$; max. shift factor $\alpha = 17$; mean $l_\Delta = 99.38$



(b) $l_p = 17$; max. shift factor $\alpha = 17$; mean $l_\Delta = 153.08$



(c) $l_p = 17$; max. shift factor $\alpha = 17$; mean $l_\Delta = 153.21$

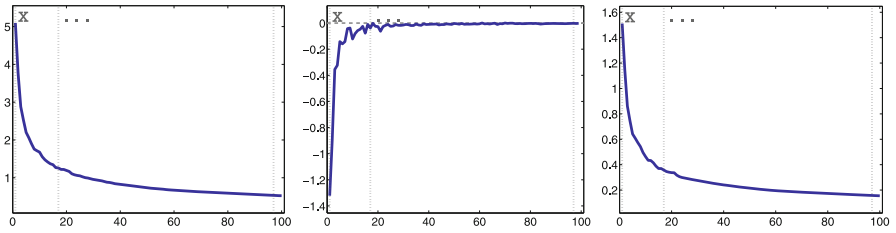


(d) $l_p = 31$; max. shift factor $\alpha = 17$; mean $l_\Delta = 278.91$

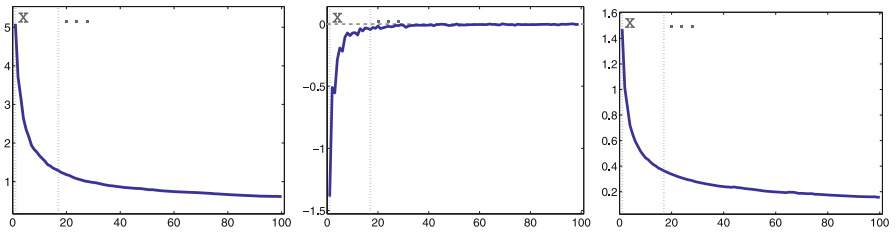
Fig. 3 Results for variable-shift corpora. x-axis: length of probe prefix in frames. y-axis: $m(x, y, \rho, k)$ (the four left panels), $m'(x, y, \rho, k)$ (the four middle panels), and $\sigma(A(x, y, \rho, k))$ (the four right panels). *Dotted vertical lines* are true segmentation boundaries

but in extremely variable form, the simple criteria we designed and evaluated did not succeed.

Based on these experiments we hypothesize that segmentation is only possible if intervals of speech are considered *in context*. What is characteristic for the methods in this section is that the prefix of the probe and the subsequence of the memory were compared without considering the context of individual frames. Given the high



(a) $l_p = 17$; $l_\Delta = 97$



(b) $l_p = 17$; max. shift factor $\alpha = 17$; mean $l_\Delta = 153.34$

Fig. 4 Results for a constant-shift corpus (above) and a variable-shift corpus (below) with a random probe. x-axis: length of probe prefix in frames. y-axis: $m(x, y, \rho, k)$ (the two left panels), $m'(x, y, \rho, k)$ (the two middle panels), and $\sigma(A(x, y, \rho, k))$ (the two right panels). *Dotted vertical lines* are true segmentation boundaries

variability of speech, very similar acoustic patterns can have very different meaning, and very dissimilar acoustic patterns can have the same meaning (cf. Peterson and Barney 1952). For example, in German single frames in the vowel [ɛ] and in the medial part of the diphthong [aɪ] are acoustically similar. Yet they correspond to different phones. A segmentation procedure that does not recognize this basic fact will posit a spurious [ɛ] in [aɪ].

In the next two sections, we will develop methods that incorporate the context of individual frames.

5 Single-Frame Baselines

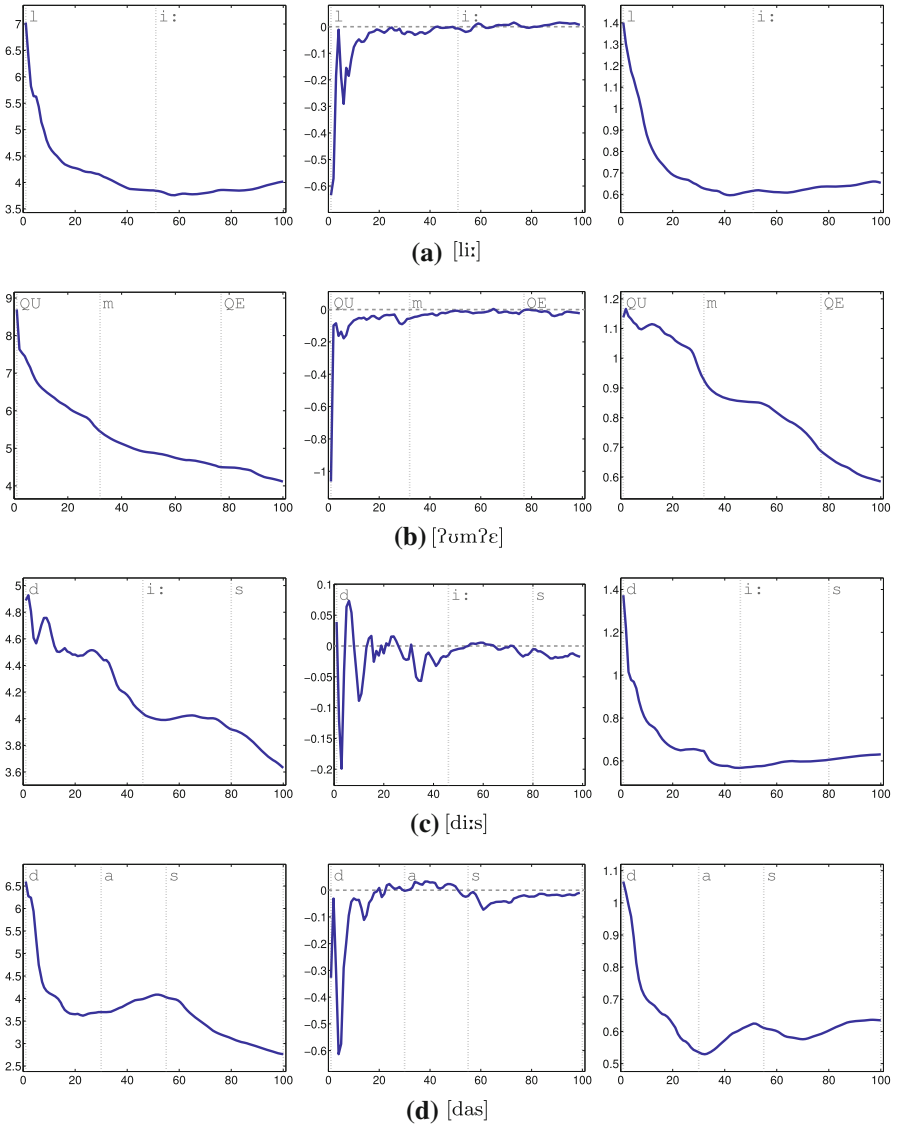
5.1 Distance of MFCC Vectors

One baseline is to define a boundariness function b_d that assigns to each frame its distance from the previous frame:

$$b_d(t) = ||\vec{v}(t) - \vec{v}(t-1)||_2$$

where $\vec{v}(t)$ is the MFCC vector of the frame at time t and $||\ldots||_2$ is Euclidean distance.

Table 1 shows results for this baseline in one experiment: two of the best (lines 1–2) and worst (lines 5–6) phones that were induced and two of intermediate quality (lines 3–4). Each induced phone is automatically labeled based on the most frequent labels that occur in it. For example, the label `ç/28-m/25-œ/18-iː/05-4` indicates that



(a) [liː]

(b) [ʔʊmʔɛ]

(c) [diːs]

(d) [das]

Fig. 5 Results for speech data. x-axis: length of probe prefix in frames. y-axis: $m(x, y, \rho, k)$ (the four left panels), $m'(x, y, \rho, k)$ (the four middle panels), and $\sigma(A(x, y, \rho, k))$ (the four right panels). *Dotted vertical lines* are true segmentation boundaries

28% of frames in the cluster had the phonetic label [ç], 25% [m], 18% [œ], and 5% [iː]. Less frequent labels are omitted. The size of the cluster (number of frames) is order of magnitude 10^4 (last digit of label).

The best clusters closely correspond to labeled phones: [t], [x]. Generally, plosives and fricatives can be most easily acquired in an unsupervised fashion using b_d. Most clusters however are mixed as the examples on lines 3–6 show. Overall purity is less

Table 1 Examples of induced phones from one experiment ($b_d, k = 100, a = 30$)

Line	Cluster
1	t/97-../01-x/00-a/00-5
2	x/90-t/02-a/02-g/01-4
3	ʃ/45-ə/25-g/05-n/05-4
4	aɪ/37-n/18-m/09-ə/06-5
5	ç/28-m/25-œ/18-iː/05-4
6	oː/25-a/21-m/19-ɐ/17-5

than 40%, indicating that less than 40% of a typical cluster corresponds to a labeled phone and the other 60% do not. These negative results are not unexpected, but they confirm that phone acquisition is hard.

5.2 Minimum Pair Distance

A variant of b_d is b_m, which—instead of the distance between two consecutive frames—computes the *minimum pair distance* (MPD), the minimum distance of any pair of two frames, one of which occurs up to $c = 10$ frames before t and one of which occurs up to c frames after t:

$$b_m(t) = \min_{1 \le i,j \le c} ||\vec{v}(t-i) - \vec{v}(t+j-1)||_2 \tag{12}$$

We call the MPD boundariness function b_m.

The intuition is that boundaries correspond to points t in time where what is left of t is different from what is right of t. Due to the high variability of speech, we do not want to rely on individual pairs of different vectors $\vec{v}$. Possible aggregate measures that are robust against producing false positive boundaries are the average distance, the median distance or the minimum. In this paper, we use the minimum as shown in Eq. (12). b_m predicts no boundary in cases where at least one vector $\vec{v}$ to the left is similar to at least one vector $\vec{v}$ to the right.

5.3 Transition Probability

We also evaluate two baseline methods that correspond to the predictability strategy (Sect. 2.1): transition probability and entropy. The boundariness function based on transition probability is defined as follows:

$$b_t(t) = -P(c(t)|c(t-1))$$

where $c(t)$ is the MFCC cluster identifier (one of $\text{MFCC}_1, \ldots, \text{MFCC}_{1000}$) of the frame at time t. We take the negative of the probability because low probabilities indicate high boundariness: we will usually posit a boundary where a very unlikely transition from one MFCC cluster to another one occurs. The transition probabilities are estimated by maximum likelihood from the corpus.

5.4 Entropy

We define boundariness based on entropy as follows:

$$b_h(t) = H(P(c|c(t-2)c(t-1)))$$

where $c(t-2)c(t-1)$ is the sequence of two MFCC clusters immediately preceding t, $P(c|c(t-2)c(t-1))$ is the probability distribution of what occurs after $c(t-2)c(t-1)$ in the corpus (estimated using maximum likelihood), and H is entropy:

$$H(p) = -\sum_i p_i \log p_i$$

Entropy measures predictability. If the underlying probability distribution P could be estimated without error, then for $b_h(t) = 0$, it would be certain which MFCC cluster will follow. For large values of $b_h(t)$, there is high uncertainty and many different MFCC clusters can follow. As with b_t, the expectation is that low predictability (in the case of b_h corresponding to high entropy) occurs at boundaries.

The four single-frame boundariness functions b_d, b_m, b_t, b_h are comprehensively evaluated in Sect. 7, together with the extended context method described next.

6 Extended Context Method

When trying to do context-dependent segmentation of speech, we are faced with a chicken-and-egg problem. Consider the context-dependent decision as to whether there is a segment boundary at point t in the signal. The context here consists of the two context intervals $i_l = (-\infty, t)$ and $i_r = (t, \infty)$ preceding and following t. Working with these potentially long intervals as units is unlikely to be helpful since, presumably, information close to t is highly relevant and information distant from t (e.g., by more than 5 s) is not relevant for the segmentation decision. This means that we need a segmentation of i_l and i_r into meaningful units, so that we can restrict the information we consider for a potential boundary at t to the rightmost segment of i_l and the leftmost segment of i_r. Unfortunately, this approach to segmentation is circular. To segment at t we must know the segmentations of i_l and i_r. The definitions of i_l and i_r in turn depend on the segmentation decision for t. The extended context method introduced in this section gets around this impasse by defining *context frames* in an adaptive manner.

As a motivating example for the extended context method consider the distinction between the vowel [ɛ] and the medial part of the diphthong [aɪ], which are (under certain conditions of stress and coarticulation) acoustically similar in German. The model needs to distinguish them to perform segmentation or it would create an artificial [ɛ] segment within [aɪ].

Context is necessary to detect the difference between these two acoustically similar sounds. The context of the medial part of [aɪ] is consistent: there is always an [a] on the left and an [ɪ] on the right. If we include this context in the representation of a

frame, effectively creating a frame-plus-context, then we can distinguish [ɛ] and the medial part of [aɪ].

Observe that the immediate neighbors of the frame do not constitute effective context. The immediate neighbors of a frame in the medial part of [aɪ] are almost identical to the frame itself and will most likely have the same cluster identifier—so they cannot be used to distinguish [aɪ] from [ɛ]. Using context at a fixed distance (of, say, $c = 10$ frames) is also suboptimal because we will skip important events, e.g., the frame at distance $c = 10$ may be "on the other side" of a plosion.

To provide consistent helpful local context for frame f, we use a threshold ρ and find the *left (resp. right) context frame*. The left (resp. right) context frame is defined as the first frame f' to the left (resp. right) with $||\vec{v}(f) - \vec{v}(f')||_2 > \rho$. This search will skip over a long interval of a steady state; on the other hand, it will pick the immediate left (or right) neighbor if significant acoustic change occurs between f and this neighbor.[4] We choose the value of ρ for which the median distance between a frame and its left/right context vector is $c = 10$ frames. This value is $\rho = 0.673$.

Our assumption here is that we need to be sensitive to acoustic changes; in particular, to skip over intervals of no change and pay attention to points of significant change.

The corpus can now be represented as a sequence of four million 36-dimensional *extended context vectors* $\vec{w}(t)$, where the first 12 components correspond to the MFCC vector of the left context frame, the second 12 components correspond to the MFCC vector of the frame itself, and the third 12 components correspond to the MFCC vector of the right context frame.

The boundariness measure we define on these vectors is b_e (e = extended context):

$$b_e(t) = \min_{1 \le i,j \le c} ||\vec{w}(t-i) - \vec{w}(t+j-1)||_2 \tag{13}$$

where, again, $c = 10$. Boundariness is computed as minimum pair distance as motivated above for b_m in Eq. (12). To speed up the computation of this boundariness measure, we use the extended context centroid vectors of each frame; this makes it possible to cache distances and speeds up computation by an order of magnitude.

7 Evaluation

Evaluation results for the methods defined in Sects. 5 and 6 are shown in Table 2 and Table 3. The two tables evaluate the five methods previously introduced: b_t, b_h, b_d, b_m and b_e. The two measures described in Sect. 3, adjusted rand index (ARI) and purity (P), are computed for two values of the number of induced phones k, $k \in \{100, 1000\}$; and for three values of the average segment length a, $a \in \{30, 50, 100\}$. Adjusted rand

[4] The sequence of left and right context frames computed this way can be nonmonotonic. For example, the left context frames of frame i and $i + 1$ might be $i - 10$ and $i - 12$, respectively, that is, left context is skipping back. To prevent this, we impose the monotonicity constraints $l'(i) \leftarrow max_{j \le i} l(j)$ and $r'(i) \leftarrow min_{j \ge i} r(j)$, where l and r are the base functions that return the indices of left and right context vectors. In what follows, we use l' and r'.

Table 2 Evaluation of the joint problem of segmentation and classification

		a	30		50		100	
		k	100	1,000	100	1,000	100	1,000
ARI	b_t		0.089*	0.023*	0.089*	0.023*	0.087*	0.023*
	b_h		0.085*	0.022*	0.074*	0.020*	0.067*	0.017*
	b_d		0.100*	0.024*	0.092*	0.023*	0.097*	0.025*
	b_m		0.115*	0.026*	0.121*	0.030*	0.116*	0.027*
	b_e		**0.163**	0.036*	0.140*	0.036*	0.114*	0.026*
P	b_t		0.352*	0.438	0.343*	0.427	0.314*	0.381*
	b_h		0.361*	0.449	0.341*	0.424	0.300*	0.364*
	b_d		0.373*	0.460	0.360*	0.442	0.320*	0.387*
	b_m		0.392	0.479	0.392	0.478	0.332*	0.403
	b_e		**0.393**	0.474	0.372*	0.456	0.329*	0.393

ARI and purity for different methods and parameters. Values are averages of 10 trials; runs that are significantly worse than b_e, $a = 30$, $k = 100$ (in bold) are marked with $*$

Table 3 Evaluation of segmentation only

		a	30		50		100	
		k	100	1,000	100	1,000	100	1,000
ARI	b_t		0.346*	0.059*	0.290*	0.052*	0.182*	0.035*
	b_h		0.331*	0.057*	0.253*	0.047*	0.148*	0.030*
	b_d		0.361*	0.061*	0.297*	0.052*	0.179*	0.036*
	b_m		0.407	0.067*	0.351*	0.061*	0.211*	0.038*
	b_e		**0.425**	0.073*	0.348*	0.064*	0.198*	0.037*
P	b_t		0.674*	0.697*	0.621*	0.645*	0.492*	0.520*
	b_h		0.677*	0.698*	0.609*	0.633*	0.467*	0.491*
	b_d		0.694*	0.719*	0.634*	0.657*	0.494*	0.514*
	b_m		0.721	0.748	0.682*	0.699*	0.510*	0.533*
	b_e		**0.726**	0.743	0.661*	0.678*	0.506*	0.525*

ARI and purity for different methods and parameters. Values are averages of 10 trials; runs that are significantly worse than b_e, $a = 30$, $k = 100$ (in bold) are marked with $*$

index and P values close to 0 indicate an essentially random clustering, i.e., phone acquisition failed. Values significantly different from 0 indicate a partially successful clustering. Using ARI, we can fairly compare clusterings of different sizes; for example, we can see in Table 2 that for $a = 30$, 100 clusters (ARI between 0.089 and 0.163) perform much better than 1,000 clusters (ARI between 0.023 and 0.036). Purity tends to increase with more clusters and does so in this case: for $a = 30$ purity values are consistently higher for $k = 1,000$ than for $k = 100$. Even though purity gives an unfair advantage to more clusters, it is easy to interpret: it is the percentage of frames that are correctly assigned to a cluster if the label of that cluster is taken to be its majority label (see definition in Sect. 3). Thus, about 35 to 40% of frames are correctly assigned for $a = 30$, $k = 100$. This is not a very good performance, but as we will see in Sect. 8 it is good enough to support successful correspondence learning.

On ARI, results for $k = 1,000$ clusters are consistently worse than for $k = 100$ clusters; this is clearly related to the fact that the true size of the phone inventory is 63 and 100 is closer to 63. However, sometimes high-cardinality clusterings capture

the structure of the data better, so the number of human-defined categories cannot be adopted without experimental verification.

The differences between the values of a are not as large, but $a = 30$ is better for most methods than $a = 50$ and $a = 100$. This indicates that it is hard to find intervals of no or little change when average interval size is 100 ms ($= 2 \cdot 50$) or 200 ms ($= 2 \cdot 100$). b_e is generally the best segmentation method for given a and k, but there are a few cases where b_m is slightly better (e.g., ARI, $k = 100$, $a = 100$ and P, $k = 100$, $a = 50$). The very best result on ARI, 0.163, is achieved by b_e for $k = 100$ and $a = 30$. Note that 60 ms ($= 2 \cdot 30$) is the typical length of a short labeled phone. This is better by about 16% than 0.140, the best result of the next-best method: b_m for $a = 50$ and $k = 100$. We will therefore use b_e as the basis for correspondence learning in Sect. 8.

In Table 2, induced phones are formed from segments represented in the space of *1,000 MFCC clusters*; in Table 3, induced phones (i.e., clusters) are formed from segments represented in the space of *63 labeled phones* to evaluate segmentation without classification (see Sect. 3). Only the evaluation presented in Table 2 is a true evaluation of the performance of the joint segmentation and classification task that we have defined phone acquisition to be.

The evaluation in Table 3 is intended to show that the variability of the speech signal is the key problem in phone acquisition. We interpret the 63-dimensional space as a version of our corpus that is constructed to be non-variable: All frames that are parts of [s] are recognized as identical. In contrast, in the 1000-dimensional space some [s]-frames are represented as identical to frames of other labeled phones, e.g., [ʃ] or [t]. As is apparent from the comparison of the two tables, phone acquisition on the realistic variable representation fares much worse than phone acquisition on the non-variable version of the corpus. For example, the best method, the extended context method b_e, achieves a high ARI of 0.425 for the non-variable corpus, but only 0.163 for the real corpus (for $a = 30$, $k = 100$; bold numbers). We view these experimental results as evidence that, in purely acoustic terms, different instances of one and the same labeled phone can be quite different and that this is the main difficulty that has to be overcome in phone acquisition.

On the fairer ARI measure, b_e is better than all other methods for $a = 30$, $k = 100$ in Table 2—each result that is statistically different and less than the b_e, $a = 30$, $k = 100$ run is marked with $*$ (two-sample t-test, samples of equal size 10, equal variance, $t > 2.262$, $p < .05$). b_e also fares well on the purity measure, but results are less clear here because of the biases inherent in purity that were discussed above.

In general, the baseline methods commonly used in the symbolic sequence approach, in particular b_t and b_h, fare poorly, indicating that the problem of learning phones is different from segmentation in the symbolic sequence approach. The probability transition method b_t does best, but even in that case performance is significantly below that of the extended context method.

8 Establishing Correspondence

As described in the introduction we assume that correspondence learning is triggered when the child recognizes that two utterances refer to the same object. Each pair of such utterances can serve as a training instance for correspondence learning. The

Correspondence-by-Segmentation Hypothesis states that, based on the segmentation learned in the experiments in Sect. 6, we can learn to recognize correspondences based on a training set of correspondence pairs. The segmentation used in this section is b_e with parameters $k = 100$ and $a = 30$.

As a formal model of correspondence learning, we adopt a simple Bayesian framework. Let $a_1 \ldots a_n$ be a word that is currently observed by the child, represented as a sequence of induced phone labels a_i, one for each frame. For example, if $a_1 \ldots a_n$ consists of two induced phones, labeled i' and i'', the first one occupying the first 8 frames, the second one the next 3 (i.e., $n = 11$), then $a_1 \ldots a_n = i'i'i'i'i'i'i'i''i''i''$.

Let $b_{k1} \ldots b_{km_k}$ represent a previously observed word w_k, represented again as a sequence of induced phone labels, one for each frame. Then we determine the previously observed word w_k that is the most likely match for $a_1 \ldots a_n$ as follows:

$$\text{argmax}_k P(b_{k1} \ldots b_{km_k} | a_1 \ldots a_n) \tag{14}$$

$$= \text{argmax}_k \frac{P(a_1 \ldots a_n | b_{k1} \ldots b_{km_k}) P(b_{k1} \ldots b_{km_k})}{P(a_1 \ldots a_n)} \tag{15}$$

$$= \text{argmax}_k P(a_1 \ldots a_n | b_{k1} \ldots b_{km_k}) P(b_{k1} \ldots b_{km_k}) \tag{16}$$

$$= \text{argmax}_k \prod_{i=1}^{n} [P(a_i | b_{kf(i,m_k)}) P(b_{kf(i,m_k)})] \tag{17}$$

where $f(i, m) = \text{argmin}_{1 \le j \le m} | j - \frac{i}{n} m |$ is the index in $b_{k1} \ldots b_{km}$ that is proportionally closest to the index i in $a_1 \ldots a_n$; that is, we assume a simple linear alignment of the two sequences.

The denominator in Eq. (15) is omitted since it is the same for all previously observed words w_k. We then assume independence of the induced phones within a sequence from each other and conditional independence of induced phones given a sequence of previously observed phones in Eq. (17). This assumption is a simplification, but it is a much better approximation for phones induced by using segmentation method b_e than for acoustic classes at a more fine-grained level (e.g., for MFCC vectors).

To summarize, Eq. (17) states that we will interpret $a_1 \ldots a_n$ as the same word as that previously stored word w_k for which the product of the conditional generation probabilities $P(a_i | b_{kf(i,m_k)})$ of the generated induced phones and the prior probabilities $P(b_{kf(i,m_k)})$ of the generating induced phones is largest. This model is applied by first estimating the parameters $P(a_i | b_{kf(i,m_k)})$ on a training set of pairs of words; and then applying it to a test set of new instances. The training set is taken from the training set of the corpus (the first 3,000,000 frames); the test instances are taken from the test set (remaining 1,300,000 frames). Correspondences of two induced phones i_f and i_g that occur in only two word pairs or fewer are discarded as noise and not used in the estimation of $P(a_i | b_{kf(i,m_k)})$.

8.1 Evaluation

Following ACORNS (ten Bosch et al. 2008), we selected a set of 13 words for the experiments. We chose those words that were nouns and occurred at least 7 times in the

training corpus of 3,000,000 frames and selected a random subset of 7 tokens from the corpus for those words with more than 7 tokens. For each word, the correspondence model was trained on all pairs of tokens of this word. Thus, the final correspondence model has been trained on 13 words, each having 7 tokens in the training corpus. We estimated the conditional probabilities using add-one smoothing:

$$P(a_i|b_j) = \frac{C(a_i b_j) + 1}{C(b_j) + 100}$$

where $C(a_i b_j)$ is the number of times that the induced phone a_i occurs in the observed word and the induced phone b_j occurs in corresponding position of the stored word and $C(b_j)$ is the total number of occurrences of the induced phone b_j. $k = 100$ is the number of induced phones. The prior $P(b_j)$ was estimated by maximum likelihood.

The model was tested by randomly selecting for each of the 13 words, one token from the test set, and comparing it as a probe against the tokens in the training set, using Eq. (17). The target with the highest conditional probability was then selected. If probe and target corresponded to the same word, then this was counted as a success, otherwise as a failure.

In 12 out of 13 cases the correct target was selected. In one case an incorrect target was selected: The training set token picked by the model for the probe "Bayern" was "Ball". Presumably, children will make some such errors in the early stages of language acquisition. We interpret a rate of 92% correct decisions as a successful demonstration that correspondence can be established based on an unsupervised segmentation method like b_e and a training set of correspondences for a small set of words.

9 Analysis and Discussion

9.1 Phone Acquisition

We performed a detailed qualitative analysis of the 100 clusters of one experiment with segmentation method b_e, $a = 30$, and $k = 100$. Although the results were clearly better than for the baselines, there was a mix of good and bad induced phones. The best cluster in terms of purity was ɔʏ/92-n/01-t/01-h/01-4. This induced phone corresponds well to the labeled phone [ɔʏ] in terms of precision (92% as indicated by the label) and recall (there was no other induced phone in which [ɔʏ] frames were among the top four most frequent). Similarly, some induced phones correspond with high precision to subclasses of [aʊ], [aː], [n] and [s]; however, in these cases recall is low: in each case there are several other induced phones that also contain many instances of these four labeled phones.

We identified several reasons for the differences between induced phones and labeled phones. In some cases, induced phones do not reflect **phonetic conventions.** For example, t/47-ə/30-x/03-ɪ/03-5 consists of instances of the transition between [t] and [ə] with weak voicing cues, which are labeled in the gold standard as either [t] or [ə]. Since there is no hard and fast criterion for setting the

boundary in this case, this induced phone does not necessarily constitute a failure of learning.

Drift

The basic assumption underlying the definition of b_e is that boundaries between phones correspond to above-average acoustic changes. But transitions between phones can be very gradual. As a result, the acoustics of what happens at time t and time $t + 50$ ms might be very different, but each point in the interval is a point of only slight change. An example is the induced phone aʊ/63-s/25-ʃ/03-z/02-4. The transition from the fricatives [s], [z], [ʃ] to the diphthong [aʊ] is gradual. As a consequence, combinations of [aʊ] with these fricatives were assigned to one cluster. One possible way to treat this, at least for certain cases, would be to augment b_e such that the threshold is sensitive to cumulative change in a constant direction, possibly weighted by the number of consecutive frames of increment/decrement.

Infrequent Phones

All native German labeled phones were dominant in at least one induced phone with three exceptions: [ʏ], [œ], [øː]. These happen to be the three German labeled phones that are least frequent in the corpus—fewer than 12,000 frames correspond to each. Clustering must give preference to frequent events and apparently the contexts of the three sounds were not consistent and distinctive enough to give rise to a cluster. A data-driven method will probably always commit a few such errors.

Fine Acoustic Distinctions

In many cases, a labeled phone was split across a number of induced phones based on acoustic distinctions. For example, two induced phones of the labeled phone [t] were t_1: t/81-ɐ/03-ɪ/02-oː/02-5 and t_2: t/73-../15-l/03-k/02-4. t_1 occurs in most cases in the middle of a segment that is labeled as [t], indicating that it corresponds to the region of plosion. In contrast, t_2 occurs either as the first induced phone in such a sequence (indicating that it corresponds to the silent occlusion phase) or at the end of the utterance when the release tapers off into silence. The latter case is also responsible for the 15% silent frames in the cluster. To make a distinction between these different manifestations of [t] could be valid for some applications like phonetic analysis, speech synthesis and speech recognition.

9.2 Correspondence Learning

There is a debate in language acquisition as to how prevalent utterance-label pairs are in the infant's input and to what extent they are crucial for language acquisition. It is true that almost any utterance can be interpreted as referring to the situation in which it is uttered and thus constitutes a potential utterance-label pair. However, all the difficulties of learning the semantics of words and utterances apply in this case.

Does the utterance "Look what Fred has found!" refer to the shape, motion, texture or color of the object that Fred is holding in his hands (cf. Quine 1960)? In this case, none of these hypotheses is correct: no words referring to shape, motion, texture or color of the object occur in the utterance. How does the child not fall into the trap of interpreting the next utterance, "It's red!", as having the same content and, consequently, into the trap of assuming that "red" and "Fred" are slightly different manifestations of the same underlying form?

Since it is difficult to assess quantitatively how many non-noisy utterance-label pairs the child has available during learning, we believe that a theory that can explain learning with a small sample of such pairs is superior to one that requires a large number. Our hypothesis is that segmentation-based correspondence learning is more robust than learning methods that try to recognize utterances as instances of words in a less structured way.

The Correspondence-by-Segmentation Hypothesis (CSH) as we stated it earlier refers to the mapping of instances of a given word from two different speakers (infant, adult 1 or adult 2) onto each other. But our experiment was conducted for a single speaker. Nevertheless, given the variability of speech, even for a single speaker, we regard the experiments we reported as strong evidence that it is possible to establish correspondence in the way hypothesized in the introduction. Still, it is necessary to show in future work that the same results can be obtained for two different speakers.

9.3 Role of Articulation

Of the correspondences in Fig. 1 the one that is most discussed and investigated in phonetics is I/AC-I/AR. This emphasis on I/AC-I/AR is due to the fact that in this case correspondence is established between very dissimilar events, e.g., for the phone [ð], (i) the directing of the tongue to move towards the teeth (a direction to a muscle) and (ii) the perception of a particular type of frication (a particular type of signal generated by the hair cells of the cochlear coil). The disparity between these two events makes it all the more surprising that every competent speaker of English has mastered this type of correspondence of production and perception.

As we discussed in Sect. 2, articulation is almost certainly an important source of information for segmentation. While there may be no purely acoustic cues that would group a sequence of three events—silence, plosion, release—into one phone, the fact that all three are caused by one continuous movement of an articulator (the tongue in case of [ð]) will bias any learning algorithm to view them as a single phone. We have not addressed the question of how and when articulation would interact with acoustics-driven segmentation in this paper even though it certainly is an important constraint on which acoustic sequences are assigned to the same category and which are ultimately viewed as different.

9.4 Innateness

It is important to stress that we call the learning procedures we developed in this paper *unsupervised* and *semi-supervised* because they do not presume that the phone

inventory is innate or that the child has access to data that would correspond to the labeled speech corpora that are used to train systems in ASR.

To the extent that we have shown that an unsupervised approach can successfully learn induced phones, the claim that phones must be innate is weakened. However, this in no way implies that innate knowledge and innate constraints do not play a crucial role in phone acquisition and correspondence learning. In general, any type of learning is impossible in the absence of some bias (such as a model or another predefined structure, (Mitchell 1980; Wolpert and Macready 1997)) or prior knowledge, which can take the form of knowing beforehand which values of a parameter are more likely than others.

To emphasize this point, we list here some of the constraints and capabilities that we have assumed and that we believe can all plausibly be argued to be innate in our account of how phone acquisition and correspondence learning unfold computationally. Ultimately, one might posit that these constraints arose as part of human evolution. All segmentation methods rely on constraints (1), (2), and (3). The extended context method in addition uses (4) and (5). The correspondence learner (Sect. 6) learns based on utterance-label pairs (6).

1. Infants can distinguish speech from non-speech.
2. MFCCs are an adequate model of what the infant perceives.
3. The infant can distinguish between smaller and greater similarity of two acoustic signals. This is the basis of clustering or vector quantization of MFCC vectors.
4. Infants can distinguish between intervals of little acoustic change and intervals of great acoustic change.
5. Infants are able to focus on the right time frame for context of a single frame. We set $c = 10$ or 20 ms. We believe that neither a time frame an order of magnitude smaller ($c = 1$) nor a time frame an order of magnitude larger ($c = 100$) would have produced usable segments or correct word correspondences.
6. In spite of the uncertainty and noisiness of the learning situation of the child there are a sufficient number of correct utterance-label pairs in the child's input to establish correspondence as outlined in Sect. 8.

Thus, the model of acquisition we propose is far from endorsing tabula rasa learning or reliance on behaviorist stimulus-response patterns. Instead, it assumes that the infant is guided by rich and complex innate knowledge. However, our claim is that this innate knowledge does not include an inventory of phones.

10 Conclusion

From the outset we stated that our fundamental assumption regarding segmentation was that it could be learned by children in an unsupervised manner and that it formed the basis for correspondence learning, a crucial component of language competence. This assumption underpins the following contributions we have attempted to make in this article: (i) we have developed a new conceptual framework for an important problem in language acquisition, the correspondence problem, (ii) we have put forward the Correspondence-by-Segmentation Hypothesis (CSH), which states that correspondence is primarily learned by first segmenting speech in an unsupervised manner and

then mapping the acoustics of two different speakers onto each other (iii) we have shown that a rudimentary segmentation of speech can be learned in an unsupervised fashion by employing our novel *extended context* method, and (iv) we were able to demonstrate that, using the previously learned segmentation, different instances of a word can be mapped onto each other with high accuracy based on a small training set of utterance-label pairs, which we feel provides corroborative evidence for our hypothesis. We also believe our demonstration of the difficulties involved when transitioning from a whole-sequence similarity method on artificial speech to natural speech illustrates the problematic nature of the task.

Although the experiments described above have, in our view, been supportive in justifying our underpinning assumption, there are nevertheless a number of areas where we feel improvements could be made. Firstly, our experiments used a unit selection speech synthesis corpus, however a child-directed (or at least "child-witnessed") speech corpus would be more appropriate. Secondly, it would be more realistic to use utterance-label pairs, in which, in addition to the targeted word, other phonetic material occurs (as is the case in ACORNS). Thirdly, correspondence learning was demonstrated on different instances of a word from the same speaker. In the future, the experiment should be repeated for different speakers. Indeed, investigations using different languages could also prove informative. Fourthly, the computational model is unrealistic in that it is batch-oriented. A cognitively more adequate incremental architecture should be pursued. Finally, human learning is typically iterative, i.e., what was learned in the previous iteration is immediately used in the next one. This aspect has not yet been incorporated into the model.

To conclude, while a number of opportunities for improvement remain, we hope nevertheless to have made a contribution to precise formal specification and computational modeling of acquisition algorithms based on real speech data. In our opinion, this approach has not been used widely enough and should, in the future, produce useful insights into how children learn language.

Acknowledgments This research was funded by Deutsche Forschungsgemeinschaft (grant SFB 732).

References

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A speech corpus for modeling language acquisition: Caregiver. In: *LREC* (pp. 1062–1068).

Aversano, G., Esposito, A., Esposito, A., & Marinaro, M. (2001). A new text-independent method for phoneme segmentation. In: *Proceedings of the 44th IEEE 2001 midwest symposium on circuits and systems, MWSCAS* (Vol. 2, pp. 516–519).

Blackburn, C. S., & Young, S. J. (1996). A self-learning speech synthesis system. In: ECSA 4th tutorial and workshop on speech production modelling (pp. 225–228).

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences, 3*(8), 294–301.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33*, 111–153.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America, 95*(3), 1570–1580.

Coen, M. H. (2006). Self-supervised acquisition of vowels in american english. In: *AAAI'06: proceedings of the 21st national conference on Artificial intelligence* (pp. 1451–1456). Menlo Park, CA: AAAI Press.

Crystal, D. (2003). *A dictionary of linguistics & phonetics*. New Jersey: Blackwell Publishing.

de Marcken, C. G. (September 1996). *Unsupervised language acquisition*. Ph.D. thesis, Massachusetts Institute of Technology.

Driesen, J., ten Bosch, L., & Van hamme, H. (2009). Adaptive non-negative matrix factorization in a computational model of language acquisition. In: *Interspeech* (pp. 1731–1734).

Fowler, C. A. (2004). Speech as a supramodal or amodal phenomenon. In G. A. Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of multisensory processes* (pp. 189–201). Cambridge, MA: MIT Press.

Frank, M. C., Goldwater, S., Mansinghka, V., Griffiths, T., & Tenenbaum, J. (2007). Modeling human performance in statistical word segmentation. In: *Proceedings of the 29th annual meeting of the cognitive science society* (pp. 281–286).

Gersho, A., & Gray, R. M. (1991). *Vector quantization and signal compression*. Berlin: Springer.

Gold, K., Scassellati, B. (2006) Audio speech segmentation without language-specific knowledge. In: *Cognitive science society* (pp. 1370–1375).

Goldinger, S. D. (1997). Words and voices—perception and production in an episodic lexicon. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 33–66). San Diego: Academic Press.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.

Goldsmith, J., & Xanthos, A. (2009). Learning phonological categories. *Language, 85*(1), 4–38.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*(1), 21–54.

Huang, X., Acero, A., & Hon, H. -W. (2001). *Spoken language processing: A guide to theory, algorithm and system development*. Englewood Cliffs: Prentice-Hall.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3*(9), 323–328.

Kuhl, P. K. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. Cohen (Eds.), *Handbook of infant perception, vol. 2* (pp. 275–382). New York: Academic Press.

Kuhl, P. K. (1988). Auditory perception and the evolution of speech. *Human Evolution, 3*(1–2), 19–43.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138–1141.

Kuhl, P. K., & Rivera-Gaxiola, M. (2008). Neural substrates of language acquisition. *Annual Review of Neuroscience, 31*, 511–534.

Lin, Y. (2004). *Learning phonetic features from waveforms* (pp. 64–70). Tech. Rep. 103, Department of Linguistics, UCLA.

Lin, Y. (2005). *Learning features and segments from waveforms: A statistical model of early phonological acquisition*. Dissertation, UCLA.

Ljolje, A., Hirschberg, J., van Santen, J. (1997). Automatic speech segmentation for concatenative inventory selection. In *Progress in speech synthesis* (pp. 305–311). Berlin: Springer.

Massaro, D. W. (2004). From multisensory integration to talking heads and language learning. In G. Calvert, C. Spence, & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 153–176). Cambridge, MA: MIT Press.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465–494.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language, 39*, 21–46.

Meltzoff, A. N., & Moore, N. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*, 75–78.

Miller, M., Wong, P., & Stoytchev, A. (2009). Unsupervised segmentation of audio speech using the voting experts algorithm. In: *AGI* (pp. 138–143).

Mitchell, T. M. (May 1980). *The need for biases in learning generalizations*. Technical report cbm-tr-117, Rutgers Computer Science Department.

Morgan, N., Bourlard, H., & Hermansky, H. (2004). Automatic speech recognition: An auditory perspective. In S. Greenberg, W. A. Ainsworth, A. N. Popper, & R. R. Fay (Eds.), *Speech processing in the auditory system* (pp. 309–338). New York: Springer.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175–184.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. L. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins Publishing Company.

Qiao, Y., Shimomura, N., & Minematsu, N. (2008). Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. In *ICASSP* (pp. 3989–3992).

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: The MIT Press.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17*(6), 405–409.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*(1), 113–146.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.

Scharenborg, O., Ernestus, M., Wan, V. (2007). Segmentation of speech: Child's play? In *Interspeech* (pp. 1953–1956).

Scharenborg, O., Wan, V., & Ernestus, M. (2010). Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *Journal of the Acoustical Society of America, 127*(2), 1084–1095.

Schweitzer, A., Braunschweiler, N., Dogil, G., Klankert, T., Möbius, B., Möhler, G., Morais, E., Säuberlich, B., & Thomae, M. (2004). Multimodal speech synthesis. In W. Wahlster (Ed.), *SmartKom: Foundations of Multimodal Dialogue Systems* (pp. 411–435). Springer, Berlin

Sharma, M., & Mammone, R. J. (1996). "Blind" speech segmentation. In *ICSLP* (pp. 1237–1240).

Slaney, M. (1998). Auditory toolbox. Online web resource, accessed: 2009-06. http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*.

Stouten, V., Demuynck, K., & Van hamme, H. (2008). Discovering phone patterns in spoken utterances by non-negative matrix factorization. *Signal Processing Letters, IEEE, 15*, 131–134.

Strehl, A. (May 2002). *Relationship-based clustering and cluster ensembles for high-dimensional data mining*. Ph.D. thesis, UTexas Austin.

ten Bosch, L., Van hamme, H., & Boves, L. (2008). Unsupervised detection of words—questioning the relevance of segmentation. In *ISCA ITRW*.

Toledano, D. T., Gómez, L. A. H., & Grande, L. V. (2003). Automatic phonetic segmentation. In *IEEE transactions on speech and audio processing* (Vol. 11, pp. 617–625).

van Segbroeck, M., & Van hamme, H. (2009). Unsupervised learning of timefrequency patches as a noise-robust representation of speech. *Speech Communication, 51*, 1124–1138.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *PNAS, 104*(33), 13273–13278.

Varadarajan, B., Khudanpu, S., & Dupoux, E. (June 2008). Unsupervised learning of acoustic sub-word units. In *ACL/HLT* (pp. 165–168).

Wade, T., Dogil, G., Schütze, H., Walsh, M., & Möbius, B. (2010). Syllable frequency effects in a context-sensitive segment production model. *Journal of Phonetics, 38*(2), 227–239.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning And Development, 1*(2), 197–234.

Wolpert, D., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.

Xanthos, A. (2003) An incremental implementation of the utterance-boundary approach to speech segmentation. In *computational linguistics in the Netherlands* (pp. 171–180).