A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events

Okko Räsänen

Department of Signal Processing and Acoustics

Aalto University, School of Electrical Engineering, Espoo, Finland

Abstract

Word segmentation from continuous speech is a difficult task that is faced by human infants when they start to learn their native language. Several studies indicate that infants might use several different cues to solve this problem, including intonation, linguistic stress, and transitional probabilities between subsequent speech sounds. In this work, a computational model for word segmentation and learning of primitive lexical items from continuous speech is presented. The model does not utilize any a priori linguistic or phonemic knowledge such as phones, phonemes or articulatory gestures, but computes transitional probabilities between atomic acoustic events in order to detect recurring patterns in speech. Experiments with the model show that word segmentation is possible without any knowledge of linguistically relevant structures, and that the learned ungrounded word models show a relatively high selectivity towards specific words or frequently co-occurring combinations of short words.

Keywords: unsupervised learning, language acquisition, word segmentation, distributional learning

1. Introduction

Segmentation of continuous speech into words is a difficult task without a priori knowledge of the auditory word forms of a language. This is due to the fact that spoken words are rarely separated by pauses or any other universal cues that would signify word boundaries equally in all languages. However, language specific cues to word boundaries exist and human infants seem to be adept in learning these cues already at a very young age, since they are able to segment word like patterns from speech at 7.5 months (Jusczyk & Aslin, 1995). Prominent and widely studied cues for word segmentation include transitional probabilities of subsequent speech sounds (Saffran, Aslin, & Newport, 1996; Saffran, Newport & Aslin, 1996), phonotactics (Jusczyk, 1993) and aspects of prosody such as intonation and linguistic stress (e.g., Cutler, 1994; Jusczyk, 1993, 1999; Thiessen & Saffran, 2004).

In a study by Saffran et al. (1996), it was pointed out that infants as young as 8 months are capable of learning transitional probabilities of subsequent syllables in an artificial language after a very brief exposure to a continuous speech stream, and that they segment words from the stream by using these probabilities. Further experiments have provided support to the idea that the learned word-like-unit structures act as lexical candidates if they are presented in a proper linguistic context (Saffran, 2001). However, the limitation of ecological validity in these studies has been in that the speech stimuli used in the experiments consisted of synthesized speech that has far less variability than real continuous speech. More recently, Pelucchi, Hay, and Saffran (2009) have shown that infants are able to use transitional probabilities in real speech spoken in a foreign language, and also by taking into account backward probabilities of speech sounds,

providing evidence that knowledge of the phonetic or syllabic system of a language is not a necessity for distributional learning.

Transitional probabilities are furthermore closely related to the concept of phonotactics, i.e., the rule system that describes which sound sequences are permissible in a language. In his work, Jusczyk (1993) has shown that 6-month-old infants do not show a preference for phonotactically legitimate sequences when compared to non-legitimate sequences, whereas infants at the age of 9 months preferred sequences that were permissible according to their native language. Although the phonotactic constraints are conceptually tied to the concept of the phoneme, the same underlying mechanism performing transition probability analysis of any general speech sounds could also explain novelty and familiarity effects in patterns of speech without the need for phonemic representation. In other words, recognizing a phonotactically legitimate phoneme sequence as familiar does not dictate that the listener has to have a fully developed categorical perception of phonemic units. This is important from the perspective of language acquisition, since it is still being debated whether the phonemic system is required for speech coding at all (see, e.g., Pisoni, 1997; Port 2007), and if it is, does it precede (Kuhl, 2004), or follow from (Werker & Curtin, 2005), lexical learning.

It has also been shown that infants might prefer other cues over transitional probabilities if they occur systematically in their native language. For example, words in English start most often with a stressed syllable and native English infants prefer to use this cue after the age of 8-9 months (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2004). However, Thiessen and Saffran (2003) have pointed out that 7-month-old infants prefer to use transitional probabilities of speech sounds in segmentation whereas 9-month-olds were relying more on syllabic stress. This is an interesting finding, since Johnson and Jusczyk (2001) have claimed that native English infants

might bootstrap their stress-based word segmentation skill from stress patterns of isolated word productions. The reason why the use of stress then emerges later than the use of transitional probabilities might be related to the issue that the bootstrapping of the stress-based segmentation with the help of isolated productions may not be as efficient at first in comparison to the analysis of recurring patterns from continuous speech. The reason why stress cues still become dominant during development is probably since they are easier to detect in English than the transitional probabilities of subsequent speech sounds, since infants under the age of one are still gradually learning the phonetic system and the CV-pairs in their native language (e.g., Werker & Tees, 1984). Linguistic stress is also much more easily generalized across speakers and situations than statistical distributions of speech sounds. This is because realizations of phonemes and words vary to a great degree depending on the speaker's characteristics, whereas cues such as timing, energy, spectral tilt, and directions of pitch changes are much more speaker invariant (see also Thiessen & Saffran, 2003).

By drawing evidence together from the distributional learning hypothesis and experimental findings, it can be hypothesized that infants might bootstrap their word segmentation process by analyzing regularly recurring stretches of acoustic signals without pre-existing phonetic knowledge (phones, syllables). These recurring segments of speech could act as preliminary lexical items that can be associated with multimodal/motor representations such as objects or actions (functional aspect) and analyzed in further detail to facilitate further speech perception (developmental aspect). By collecting and analyzing the preliminary lexical items, infants are able to detect language specific systematical properties of words such as trochaic stress in English, and by coupling speech perception to their own articulatory productions, they start to learn sub-word structures such as syllables and phones (see, e.g., PRIMIR-theory of

language acquisition by Werker & Curtin, 2005). Although the preliminary lexical representations are highly dependent on detailed acoustic features, and are therefore speaker and speaking style dependent (Houston & Jusczyk, 2000; Bortfeld & Morgan, in press; Singh, White, & Morgan, 2008), the gradually developing heightened sensitivity to native contrasts and increase in phonemic awareness (White & Morgan, 2008) may facilitate further word learning and help to generalize across speakers (see also discussion in Swingley, 2005). The lack of well-developed speaker independent models of phonemic categories in early lexical acquisition would also explain why even 14-month-old infants have difficulties in distinguishing minimal pair words such as "*bih*" and "*dih*" from each other when spoken by the same person (e.g., Stager & Werker, 1997), but succeed in the task when notable variation is introduced to the spoken words (Rost & McMurray, 2009). In general, it seems that variability enables statistical learning of exemplar "clusters" that reveal structural differences between the words, whereas sufficiently detailed awareness of phonemic distinctions simply does not exist at early stages of development or is overrun by lexical competition (see Rost & McMurray, 2009, and references therein).

The plausibility of the above hypothesis as a mechanism for the bootstrapping of infant speech perception would be supported considerably if a computational mechanism demonstrating such processing existed. As for computational models of word segmentation from continuous speech, in Räsänen, Laine, and Altosaar (2008) and Räsänen, Laine, and Altosaar (2009a), it has been shown that automatic word segmentation based on transitional probabilities of atomic acoustic events is possible in a weakly supervised learning framework where a learning agent receives multimodal support from a visual scene. By associating recurring segments of speech signals to objects in the visual scene through cross-situational learning, the learning agent learned to recognize keywords from the incoming utterances. However, this learning paradigm

did not lead to the learning of words that were not systematically related to objects in the surrounding visual environment. Instead, only the keywords that were present as both audio and as visual categories were learned and segmented properly.

In the current work, a computational model for purely unsupervised acquisition of acoustic word form representations is proposed. Instead of multimodal support or assuming any a priori phonetic or linguistic knowledge such as phones, phonemes, or words, the processing starts with acoustic signals taken from a speech corpus containing child directed speech. The proposed algorithm tracks the transitional probabilities of atomic speech sounds in order to detect recurring patterns, and builds models for these detected patterns. As more speech is perceived, these learned models inherently segment the new utterances into words. The overall results show that it is possible to learn ungrounded models for word-like acoustic units from continuous speech without learning the phonetic system of a language, and without support from contextual information such as different modalities.

1.1 Related work

1.1.1 Phonemic and syllabic algorithms. Most of the existing models of word segmentation assume some sort of pre-existing linguistic knowledge, often in the form of phonemic transcription of speech or pre-defined phonetic features created from a text corpus. In the work of de Marcken (1994), the Minimum-description length (MDL) principle (Rissanen, 1978) was used to construct an MDL-optimal grammar from a corpus with a finite alphabet. The algorithm attempts to find such a dictionary of words that maximizes the likelihood of the word dictionary given the training data, and simultaneously minimizes the number of bits required to describe the dictionary and the corpus coded using the entries in the dictionary. The algorithm was mostly designed for the analysis of text and phonetic transcriptions. In contrast to the

algorithm proposed in this work, de Marcken assumed a phone-level transcription and knowledge about phonetic features provided either by manual annotation or a supervised Hidden Markov model–based speech recognizer for learning word patterns from continuous speech.

Phonetic transcription was also assumed as a starting point in the work of Cairns, Shillcock, Chater, and Levy (1994), who trained a connectionist neural network for the word segmentation task. As an input to the network, they used phonological features motivated by Government Psychology (Shillcock, Lindsey, Levy, & Chater, 1992) that were derived from the phonetic transcription of speech. Their algorithm performed only slightly above chance level, leading the authors to conclude that phonotactics provide a fairly weak source of information for bootstrapping of segmentation (Cairns et al., 1994). Other algorithms tested on phonetic transcription data include a connectionist model by Christiansen, Allen, and Seidenberg (1998), a further MDL-based algorithm by Brent and Cartwright (1996), the somewhat similar but incremental probabilistic approach known as the Model-based dynamic programming (MBDP-1) algorithm by Brent (1999a), and the algorithm by Venkataraman (2001). All of these models have shown good performance on the segmentation of phonetic transcripts, and in contrast to the results of Cairns et al. (1994), give strong support to the hypothesis that the unsupervised word segmentation based on transitional probabilities of speech sounds is possible, especially if phoneme level knowledge is inherent to the system. However, none of these approaches have been primarily designed, nor have been evaluated, using continuous speech without a priori linguistic assumptions.

Instead of using purely linguistic input, multimodal pattern discovery has been also studied in the context of language acquisition. A model of early language acquisition by Yu, Ballard, and Aslin (2005) combines visual and auditory processing by using images and speech

related to the images as input to the system. In their experiments, the incoming speech was transcribed into phoneme sequences using pre-trained recurrent neural networks (RNN). Attentional modeling was embedded in the visual processing, where visual scenes were segmented based on gaze information and represented by a collection of visual features. Their algorithm showed good word segmentation accuracy by using dynamic programming to detect recurring subsequences from the phoneme sequences while processing was being modulated by attentional factors (Yu et al., 2005). However, linguistic assumptions were still embedded to the speech-to-phonemes conversion in the system.

Finally, Swingley (2005) has demonstrated the segmentation of a language corpus using co-occurrence probabilities of syllables instead of phonemes. In his approach, Swingley constructed a mutual information[1] (MI) ranking for bisyllables in the corpus. In addition, frequencies of mono-, bi- and trisyllables were computed. Then the syllable sequences high in both mutual information and frequency were considered as word hypotheses. This led to a distinction between words and non-words in the upper end of the frequency+MI continuum, especially in the case of shorter words. However, there also was a notable amount of erroneous word hypotheses (Swingley, 2005).

1.1.2 Learning of speech sound categories. The numerous examples above indicate that the word segmentation task using linguistic input is possible with the help of phonemic identities of speech sounds. However, the limitation with the above approaches from the perspective of infant word segmentation is that the assumption of the existence of phonemic coding of speech is

[1] Swingley (2005) used mutual information $MI_{AB} = \log_2[P(AB)/(P(A)P(B)]$ of consequent syllables A and B in his approach. However, this can be considered as modeling of transitional probabilities between syllables, simply described from a slightly different perspective. Namely, the approach measured the statistical independence of the two subsequent syllables.

questionable. Although the idea that learning of phonemic categories precedes lexical learning is widely cultivated (e.g., NLM-e theory by Kuhl et al., 2008), there are also contrasting views. For example, the PRIMIR-theory of language acquisition by Werker & Curtin (2005) hypothesizes that phonemic knowledge emerges from the overlap between word forms stored in the learner's memory, i.e., a large number of words has to be learned before any kind of phonemic encoding can take place. Moreover, the entire existence or need for phonemic representation in speech perception has been questioned (Pisoni, 1997; Warren 2000; Port, 2007). Pisoni (1997) and Port (2007) have argued strongly against an abstract symbolic (phonemic) segmental representation of speech and give evidence that listeners do not only encode abstract linguistic messages, but also very specific acoustic details that are normally considered as indexical (non-linguistic) information. For example, listening tests indicate that listeners can exploit speaker specific acoustic properties, leading to enhanced word intelligibility for familiar versus novel speakers (see Pisoni, 1997, and references therein). This suggests that the perceived speech is not only stored as abstract messages such as sequences of phonemes, but also as detailed acoustic episodes, and that these episodic representations interact with the perception of speech. Port (2007) also notes that the phoneme-like abstractions can always be computed from overlapping properties of stored and detailed sensory episodes (cf. PRIMIR theory of language acquisition), but the abstractions alone cannot explain the phenomena observed in the perception of speech.

Even if the existence of a phonemic representation of speech in the human brain is assumed, the learning of phonemic categories directly from continuous speech faces enormous difficulties: before learning phonemes, the learner needs to segment corresponding phones out of continuous speech and find a correct mapping between different allophones across different contexts and speakers. Since realizations of different phonemes can be ambiguous and overlap

largely in the acoustic space, the distributional learning of phonemic categories may be impossible without additional constraints. Such constraints could emerge from the lexical level and other modalities, but then the supporting lexical level cannot be based on phonemic representations before learning of the phonemes. As the definition of a phoneme states that it is a minimal unit that causes contrast between two words, it is difficult to envision the emergence of phonemic knowledge in the absence of lexical representations. Although Kuhl (1986, 2004) has introduced the idea of basic cuts as a preliminary mechanism for segmenting speech sounds out of continuous speech, currently there are no unsupervised computational algorithms that could demonstrate successful segmentation in a way that would enable discovery of all phone boundaries without a large amount of hypothesized segment boundary insertions (see, e.g., Scharenborg, Ernestus, & Wan, 2007; Räsänen, Laine, & Altosaar, in press). This is also supported by the finding that not a single computational model has so far been able to demonstrate correct unsupervised acquisition of phonemic categories in the absence of top-down support, multimodal input, or directly from continuous speech. However, promising results have been obtained using greatly simplified experimental settings.

In the work of McMurray, Aslin and Toscano (2009) and Toscano & MacMurray (2010) a Gaussian mixture model utilizing a special mechanism for mixture component competition demonstrated successful unsupervised acquisition of voice onset time (VOT) categories on VOT data measured from speech. In addition, Vallabha, McClelland, Pons, Werker, and Amano (2007) have shown that a small number of distinctive phonetic categories can also be learned with a similar algorithm utilizing competition between multivariate Gaussian distributions and expectation maximization (EM) training, and also with their own incremental variant called Topographic Online Mixture Estimation (TOME). They used data from the first two formant

frequencies and durations of phones in the acquisition of English /I, i, ɛ, e/ and Japanese /i, i:, e, e:/ vowels extracted from monosyllabic words.

Coen (2006) has demonstrated that the acquisition of the number of vowel categories and their boundaries in the F1/F2 space is possible by combining the vowel formant information with lip data in a multimodal clustering process. Since acoustically ambiguous sounds are often unambiguous in terms of visual features, multimodality helps to differentiate between acoustically similar but phonetically distinct classes ("*intersensory disambiguation*"). Coen demonstrated the performance of the algorithm using a fixed acoustical context in the production of the vowels (each vowel was spoken between [h] and [d], e.g., /ae/: "had"; Coen, 2006).

Lexical feedback has also been utilized. Feldman, Griffiths, and Morgan (2009) have shown that the correct mapping from vowel formant frequencies to phonemic categories of English can be learned using a Bayesian classifier, but only if there are additional constraints from the lexical level that is being learned simultaneously with the phonemic categories. The data consisted of combinations of vowels represented by their formant frequencies that were taken from the data of Hillenbrand, Getty, Clark, and Wheeler (1995).

Both Coen's and Feldman et al.'s work demonstrate well that the phonemic learning process is greatly facilitated if additional constraints can be introduced by either feedback from the lexical level or by utilizing multimodal information sources. However, none of the above models were tested on continuous speech, where even the segmentation to phone-like units is difficult in a purely bottom-up manner (Räsänen et al., in press; Scharenborg et al., 2007), or on a full spectrum of phonemic categories. These constraints effectively limit the overall variability and overlap of phonemic categories, making generalization of the findings difficult to real continuous speech and the general speaker population.

By looking at the theoretical considerations and experimental evidence discussed so far, it seems that if phonemic representations of speech exist, then their contents must be abstracted from (and are therefore causally dependent on) lower level lexical representations with a much greater amount of sensory (and possibly articulatory) detail (as in PRIMIR; Werker and Curtin, 2005) or they can be partially a product of literacy training, as suggested by Port (2007). Acquisition of internal representations of speech is not therefore necessarily tied to the concept of a phoneme, but one can understand and produce spoken language by simply detecting and reusing acoustic patterns with sufficient similarity.

1.1.3 Towards segmentation from real speech. The similarity principle of long acoustic patterns has been previously utilized in three computational models of unsupervised word learning from continuous speech.

The PERUSE algorithm by Oates (2002) discovers frequently recurring patterns from multivariate time-series (such as automatically extracted speech features) by modeling patterns as sequences of observations with mean and variance of possible observation values defined for each temporal location in the sequence. The algorithm starts by performing a global exhaustive search over all available speech data in order to find a model that has the highest likelihood when trained with a sub-segment of length L and its N best-matching occurrences in the data (L and N are used set parameters). Then the pattern length L is increased and a statistical test is performed for the new model in order to determine whether the pattern likelihood has dropped significantly. Finally, the number of tokens used to train the model is increased from N in an incremental manner and statistical testing is again used as a stopping criterion. Oates has demonstrated the performance of the algorithm in word learning from English, German and Mandarin speech, where it successfully detected more than 65% of frequent words used by a single speaker (per

language). Oates has also represented a framework that allows grounding of the detected word forms to contextual sensory data collected by a robot (Oates, 2001). As a drawback, the PERUSE algorithm assumes that all possible speech data are available to the system when the learning begins (i.e. batch processing), and this data set has to be analyzed iteratively several times in order to converge to a set of word models. From the computational point of view, this also makes the algorithm extremely slow for large data sets. In addition, each word has to occur several times in the data before a representation can emerge for it. The author has acknowledged that iterative batch processing is an unreasonable requirement for a computational agent that should support continuous long-term language acquisition (Oates, 2001). Still, the PERUSE shows that an unsupervised system can converge to a set of pattern models learned from real speech, and that at least some of the models match very well to word-like units (detailed analysis of the results in addition to word detection rate is not reported).

In the work of Park and Glass (2005, 2006), a dynamic time-warping (DTW) algorithm was used to find similar stretches of speech from an MIT lecture corpus. Acoustically similar segments were then linked to each other through graph clustering. Their results showed successful detection and clustering for a number of words occurring several times in the speech material.

Finally, the cognitively inspired system by Aimetti (2009) performs unsupervised acquisition of word models using a dynamic programming (DP) based algorithm called DPn-gram for detection of recurring units between acoustic episodes. Aimetti has demonstrated the performance of the algorithm in an ecologically plausible multimodal learning task where the learner has to first learn lexical candidates from a child directed speech stream and then ground these items to co-occurring visual information. After training, a successful mapping from

acoustics to co-occurring visual categories was obtained. Both of the proposed DP-based systems (Glass and Park, 2005, 2006; Aimetti, 2009) assume that the learner is capable of storing all perceived auditory signals as detailed spectrotemporal trajectories, or acoustic episodes. Then the spectral distances between these episodes are computed pair wise in order to detect similar stretches of speech that are then clustered together as lexical or sub-word unit candidates. In other words, they take an episodic exemplar-based approach to the problem of word learning.

To our knowledge, the algorithm proposed in this paper is the first computational model that demonstrates unsupervised and incremental word segmentation from continuous speech utilizing transitional probabilities of speech sounds without a priori linguistic assumptions. We simply assume that the learner is able to 1) perceive speech sounds on a Mel-scale with a time-resolution of 10 ms, 2) group acoustically similar sounds into discrete classes based on an Euclidean spectral distance measure, and 3) track transition probabilities of discrete speech sound events at different temporal distances up to a few hundred milliseconds. When compared to the PERUSE and DTW-based models, the proposed approach does not store all episodic representations in full detail for infinite duration, but only incrementally stores statistical dependencies between atomic acoustic units in the context of each word model, and uses these statistics to recognize new inputs. The approach is a hybrid between the classical division of exemplar and prototype models. No single word realization is stored in detail, but several different realizations can still have the same probability of belonging to a specific word model due to parallel modeling of several spectrotemporal trajectories.

The organization of the remaining paper is as follows: in the next section, the speech material used in the experiments is introduced. In section 3, the learning algorithm is presented in detail. Section 4 shows results from the unsupervised word learning experiments using speech

from a single caregiver and then from four caregivers. Finally, in the last section, implications of the results are discussed and conclusions drawn.

2. Material

Speech material was taken from the child-directed speech corpus CAREGIVER (Altosaar et al., 2010). In these experiments, an English portion of the corpus was used. Since the original use of the corpus was to study the learning of keywords from speech, the speech material has been designed so that in addition to a set of carrier sentences, there are 1-4 keywords (nouns, adjectives and proper names) embedded in each utterance. In total, there are 50 different keywords. The keyword selection was largely based on the on-line available MacArthur-Bates Communicative Development Inventory (Fenson et al., 2003). However, unlike in real speech, the keywords are statistically balanced over the entire corpus by generating the utterances with a finite state grammar without semantic constraints. This was done to avoid strong statistical word-to-word dependencies, leading to semantically incoherent but grammatically correct productions. In this work no differentiation between keywords and other words was made, yielding a vocabulary of 80 different words as additional verbs, prepositions, articles, etc. Inflections were treated as separate words. Due to the simplicity of the vocabulary, this caused only a small number of words to split into a basic form and the third person present tense: “give/gives”, “like/likes”, “see/sees”, and “take/takes”. A full list of words in the vocabulary can be found in Appendix A.

As for talkers, this section of the CAREGIVER corpus contains continuous English speech spoken by ten different individuals. The four main talkers (the “caregivers”; two males, two females) each speak 2397 utterances including two repetitions of each utterance and two

isolated productions of each keyword. There are also six additional talkers each speaking 600 utterances, including one isolated production of each keyword. Since the idea was to study word segmentation from continuous child-directed speech spoken by a caregiver or a small number of caregivers, none of the additional talkers or isolated keyword productions were used in these experiments. This yielded a total of 2144 utterances per talker. The mean length of an utterance was 5.96 words. Speaking style was elicited child-directed speech, i.e., the talkers (who were parents) were asked to speak as if they were speaking to their child who's age was less than one year. The recordings were performed in an anechoic chamber at a sample rate of 44.1 kHz using a high quality condenser microphone. For the present experiments, the signals were downsampled to 16 kHz.

For single talker experiments, 1800 utterances were used in the training phase by concatenating them into one long signal. Each utterance was padded with 1.5 seconds of silence. The remaining 344 utterances were used for testing, and were fed to the system for segmentation sequentially. For multiple talker experiments, 4*1800 = 7200 utterances from all four main talkers were used for training and the remaining 4*344 = 1376 utterances were used for testing. In all experiments, the training set and the test set were randomly chosen. In all but the talker blocked multiple caregiver case, the order of the signals in the concatenated training signal was also randomized. The results for the evaluation of different parameter settings were always computed using an identical randomized order to ensure that ordering did not influence learning measurement.

3. Methods

This section describes the details of the computational model used for unsupervised word learning. At first, it is important to note that the term *word* will here refer to the longest structures in the speech signal that recur systematically across different speech acts, or utterances. The algorithm itself is fully unaware of the concept of word, and since the speech is not accompanied with any other categorical information that would enable grounding of the word forms, the detected structures do not carry any meaning. The purpose is simply to study what type of structures can be learned based on transitional probabilities between atomic acoustic events, and how these learned structures relate to the concept of words defined by an experienced language user. The approach does not assume any innate linguistic structures, and all processing is purely bottom-up. What is assumed by the algorithm operator is the amount of signal (in milliseconds) analyzed together in order to make an old/novel distinction for the corresponding part of the signal, the maximum temporal distance up to which statistical dependencies between acoustic events should be taken into account in the transitional probability modeling, and the amount of temporal smoothing applied to the activations of internal representations.

There are several steps in the process of modeling discovery of word segments from raw speech signals. First, incoming speech is transformed into frames of features that describe the spectral content of the signal in a compact manner. In order to learn recurring structures, or *patterns*, from speech, the continuous domain feature representations are further transformed into a series of discrete events using vector quantization of the feature vectors. Finally, a method for learning recurring patterns from speech by using transitional probabilities is applied. These stages will be discussed in the following sub-sections.

3.1 Pre-processing

The aim of preprocessing is to transform the raw speech waveform into a series of discrete acoustic events so that recurring structures, or *patterns*, can be detected. In order to represent the spectral content of speech in a compact manner, Mel-frequency cepstral coefficients (MFCCs; Davis & Mermelstein, 1980; see Appendix B for details) are utilized. MFCCs are widely used as features in speech technology, e.g., in automatic speech and speaker recognition. This is since MFCCs are readily computable, and they describe relevant aspects of the speech signal using only a small number of coefficients mimicking the spectral resolution of the human ear. In this work, 12 MFCC coefficients were extracted with a window of length 32 ms and window step size of 10 ms, i.e., the signal was described with 12-dimensional feature vectors occurring every 10 ms. The given window size and step size provide a reasonable compromise in spectrotemporal accuracy and the overall amount of feature data.

MFCC vectors as such are not suitable for analysis of transitional probabilities since each coefficient in a MFCC vector takes its value from the continuous cepstral domain. Therefore, vector quantization (VQ) was applied to the MFCCs in order to describe each signal frame with one discrete label from a finite alphabet. A randomly chosen subset of MFCC vectors (10000 frames) from the training material was used as input to a fully unsupervised k-means clustering (MacQueen, 1967) that produced a codebook of N_A discrete categories, where N_A is a user defined parameter. Next, all of the MFCCs extracted from the speech material were vector quantized by finding the nearest cluster center in the codebook in terms of Euclidean distance. Finally, the MFCC vector was replaced by the integer label corresponding to the cluster. Now the speech signal is represented by a sequence of discrete VQ labels $X = \{a_1,a_2,\ldots,a_L\}$ where $a_i \in \{1,\ldots,N_A\}$, with one VQ label occurring every 10 ms.

Although the k-means algorithm in its basic form operates in a batch mode, it can be replaced by any incremental clustering method for increased ecological plausibility. For example, the OME algorithm (Vallabha et al. 2007; Lake, Vallabha, & McClelland, 2009) or self-learning vector quantization (SLVQ; Räsänen, Laine, & Altosaar, 2009b) can both be applied for learning of categorical speech sound classification in an incremental manner and without defining the number of acoustic categories in advance. However, none of the above algorithms learn proper *phonemic* categories from continuous speech, but simply map incoming continuous spectral vectors into a finite number of acoustic classes.

3.2 Transitional probability analysis

The task of the transition probability analysis is to 1) discover recurring, temporally distributed, patterns from speech, and 2) build models for these patterns that enable recognition of similar patterns in future input. This section describes how models of patterns can be learned automatically and incrementally from speech using the transitional probability framework. It should be emphasized that the term *model* here refers to a structural description of an unspecified but significant pattern that recurs in the data, but in practice the learned models will mostly correspond to word-like units. This is partially explained by the temporal parameters of the algorithm, but also because the words (or combinations of often co-occurring short words) are the largest structures in speech that recur several times in the corpus in a relatively coherent form.

The starting point of the process is the discrete VQ sequence X produced by the preprocessing of the speech. In its basic form, the transitional probability (TP) of an element pair could be defined as $P(a_1,a_2) = F(a_1,a_2)/F(a_1)$ where $F(a_1,a_2)$ is the frequency of an ordered pair a_1,a_2 and $F(a_1)$ is the frequency of a_1 alone over the entire data set (Saffran et al., 1996).

However, since we also want to build *models* for recurring patterns, a mechanism is needed that differentiates between different models that are learned from the data. Instead of computing a global probability of a transition in order to detect word boundaries, we are interested in how probable a transition is in case of a specific model, and whether there is a previously learned model explaining the current transitions occurring in the signal.

Another challenge in using the above definition of probability stems from the complexity of real continuous speech that causes the VQ data to be extremely noisy and variable. Two realizations of a same word, even spoken by the same talker, are never the same in terms of spectrotemporal trajectories. This means that the words are also represented by more or less different VQ sequences. Moreover, the blind pre-processing is not temporally synchronized to any of the linguistic structures existing in speech, and therefore linguistic units such as syllables and words can have a very different number of VQ labels in different realizations given their normal temporal variation. The lack of synchrony also means that the feature-extraction process extracts spectral information from transition points between subsequent phones, leading to a large number of poorly defined spectral representations since the computation of the Fourier-spectrum assumes that the spectrum is stationary inside the analysis window. The variability also makes the use of standard n-gram based approaches infeasible due to the combinatorial explosion for units longer than a few tens of milliseconds. Therefore, a more robust approach to transitional probability model is required and is introduced below. The method has similarities with Hidden-Markov Models (HMMs) that are widely applied in machine learning, but does not require a priori definition of the number of models or states per model, iterative training, nor bootstrapping with annotated training data. In addition, the proposed model is able to capture

long-range temporal dependencies without making the Markov assumption (independence of subsequent states) that does not hold for speech VQ data.

3.2.1 Learning of models. The system starts out void of any models for patterns. When the first speech input X arrives, transitions in the first L_r elements in a sub-sequence X_T are used to create the first model c_1. In the model, transition probabilities between elements a_i and a_j are modeled in parallel for a number of lags $\boldsymbol{k} = \{k_1,k_2,\ldots k_K\}$. In other words, transitions are not only modeled from $X[t]$ to $X[t+1]$, but for all distances $X[t]$ to $X[t+k]$ for all $\boldsymbol{k}$.

Instead of computing the global joint probabilities for element pairs, a transition probability matrix is computed according to Eq. (1) from the frequencies of transitions. The obtained right stochastic matrices describe the future distributions[2] of labels a at $X[t+k]$ given the $X[t]$ and the model c. In the equation, $F_c(a_i,a_j|k)$ denotes the frequency of transitions from a_i to a_j at lag k and for the case of model c.

$$P_c^{\mathrm{S}}(a_j \mid a_i,k) = \frac{F_c(a_i,a_j \mid k)}{\sum_{j=1}^{N_{\mathrm{A}}} F_c(a_i,a_j \mid k)} \qquad (1)$$

The analysis window of length L_{r} is then shifted L_{s} frames and the existing models are used to recognize the new sub-sequence X_{T+1} (note that during the second window position $T+1$ there is only one model from the first window position). First, activation $A_c(t)$ of each model c at each moment of time t is computed by calculating the mean of the transition probabilities over all different lags:

[2] This formulation is similar to the estimation of the **Q** matrix in the generalized mixture transition distribution (MTDg) by Raftery (1985). The estimation of probabilities directly from transition frequencies in the training data has been proposed by Ching, Fung, and Ng (2004). However, estimation of lag specific weights ϕ_k characteristic to MTD models is not performed due to the incremental one-pass nature of the learning algorithm used in this work.

$$A_c(t) = \frac{1}{K}\sum_{k=1}^{K} P_c^S(X[t] \mid X[t-k], k) \quad (2)$$

The cumulative activation of each model is then calculated over the entire window and normalized by the window length:

$$A_c^{cum}(T) = \frac{1}{L_r}\sum_{x=T}^{T+L_r-1} A_c(t[x]) \quad (3)$$

where T denotes the window position. Now if activation A_c^{cum} of the most activated model c_M exceeds a pre-defined familiarity threshold t_r, the transition frequencies in the current window of analysis X_{T+1} are used to update the statistics of the model c_M according to Eq. (1). In other words, if a sufficiently familiar pattern (e.g., a word or syllable) is detected in terms of previously learned models, this realization of the pattern is used to update the existing model of the pattern. On the other hand, if no model achieves a sufficiently high activation, a new model c_N is created from X_{T+1} using Eq. (1). This process is repeated for the entire training data set, producing a set of models that incrementally increase their selectivity towards specific structures in the speech signal. The number and properties of the learned models depend mostly on window length L_r, window shift L_s, and the activation threshold t_r.

The following pseudo-code illustrates the learning process:

```
1) extract sub-sequence X_T of length L_r from the current window position
2) recognize X_T using the existing models
    if highest_activation > threshold t_r
      update best matching model c_M using transitions in X_T
      shift analysis window L_S steps
    else
      create a new model c_N using transitions in X_T
      shift analysis window L_r-L_S steps
    end
```

```
3) repeat steps 1-2 until all input is processed
```

In practice, if the window shift L_s is smaller than the window length L_r, the model update using X_T has to be performed only after the window has moved to X_{T+Lr}. Otherwise the previously updated model will always see a part of the signal that has already been used to train the model. Also, when a new model is created, the window is moved forward L_r-L_s frames in order to avoid the creation of several new models for a novel pattern of approximate length L_r.

3.2.2 Segmentation using the models. The models obtained according to the procedure described above can be already used to segment speech signals into stretches of model activations. However, by introducing an additional normalization procedure, the classification of novel input into existing categories becomes more efficient:

$$P_c(a_j \mid a_i,k) = \frac{P_c^{\mathrm{S}}(a_j \mid a_i,k)}{\sum_{m=1}^{N_C} P_m^{\mathrm{S}}(a_j \mid a_i,k)} - \frac{1}{N_{\mathrm{C}}} \qquad (4)$$

What takes place in Eq. (4) is that all learned models are contrasted against each other by diving a probability of a transition in a given model with the sum of the probabilities of the same transition across all known models. In addition, $1/N_C$, where N_C is the total number of models, is subtracted from the transition probability. This forces the sum of activation across all models to be zero at all times, and if a transition is equally frequent for all models, it does not have an impact on the overall activation. The normalization in Eq. (4) can be considered as a forced choice task: given a transition, what is the relative likelihood of each model if one of them has to be selected?

Now, when faced with a novel utterance X, the algorithm computes activation $A_c(t)$ for each model c at each moment of time t using Eq. (2), but now using P instead of P^{S}:

$$A_c(t) = \frac{1}{K}\sum_{k=1}^{K} P_c(X[t] \mid X[t-k], k) \quad (5)$$

This produces a temporally local estimate of model activity. Since the aim is to study temporally larger structures than single frames and the activation values have high variance even for familiar patterns due to complexity of the acoustic signals, it is useful to smooth activation over time. This can be done by low-pass filtering the activity curves. In this study, a non-weighted simple moving average (SMA) filter of length 480 ms (48 frames) was applied to the activity curves, since it was found to lead to reasonable results:

$$\widehat{A}_c(t) = \frac{A_c(t-47) + A_c(t-46) + ... + A_c(t)}{48} \quad (6)$$

In other words, the activity of a model at time *t* depends also on the activity level of the same model during last few hundred milliseconds. If a model sees a familiar pattern, its activation will rise notably above zero and stay there for the duration of the familiar event. If a model no longer receives activation from the transition probability analysis, its activation decays to zero and other competing models will become more active. The winning model for each moment of time is the one with the highest activation level (cf. the TRACE model of speech perception by McClelland & Elman, 1986).

In the following experiments, word segmentation was performed by first computing model specific activations for each moment of time in novel (untrained) utterances and then by placing word segment boundaries at locations where the most activated model changes from one to another. In the remainder of this article, the term *model activation* will refer to a time segment that starts from the point where the word model under consideration exceeds all other models in activation level, and ends at a point where another model becomes more active.

3.3 Evaluation measures

The temporal accuracy of word segmentation was evaluated by comparing the locations of detected word boundaries to the boundaries produced by automatic Hidden Markov model (HMM)-based forced alignment segmentation. Although automatic, the quality of HMM–based segmentation can be considered comparable to the quality of manual word level annotation (see, e.g., Toledano, Hernández Gómez, & Villarubia Grande, 2003). The quality of the HMM segmentation was also verified manually for several signals. During evaluation, the neighborhood of each reference boundary was searched for boundaries produced by the learning algorithm and the distance to the nearest algorithm boundary was measured. The standard deviation σ of distances over all reference boundaries was used as the quality measure of the segmentation. Additionally, the mean number of insertions per annotated word was computed to ensure that an apparent increase in segmentation accuracy was not achieved simply by an increase in the number of segment boundaries. Finally, the proportion of correctly detected word boundaries was measured by searching the neighborhood of each reference boundary for a boundary produced by the algorithm. If the nearest boundary was located inside a window of maximum allowed deviation, it was considered as correctly detected. Correct detections were computed over a range of maximum allowed deviations.

The ratio of the number of detected words to the number of annotated words was used to measure what proportion of the words in the reference was detected. This measure, called *lexical coverage C,* was computed simply as $C = 100*N_d/N_w$ where N_d is the number of detected segments exceeding 150 ms in length (user set minimum word duration) and N_w is the total number of words in the annotation. Ideally, lexical coverage should either increase as more

speech is perceived, or stay stable while model selectivity increases, both indicating increased modeling accuracy of recurrent long patterns in the input.

The contents and quality of the learned models were also analyzed. Model selectivity was measured as the entropy of the distribution of word classes represented by a model. Entropy was chosen as the selectivity measure because it does not only indicate the proportion of the most dominant word of the model, but considers the overall distribution of all categories for the given model (see Huang, 2008). As the number of different words sharing a single model decreases, so does its entropy. In order to compute the entropy of a model c, the temporal segments of speech were detected where the model c was most active. Only segments exceeding 150 ms in length were included in further analysis. These segments were compared to the underlying word-level annotation in order to obtain a distribution $P_c(\alpha)$ of underlying words $\boldsymbol{\alpha} = \{\alpha_1,\alpha_2,\ldots,\alpha_N\}$ for the model c, i.e., the number of frames n_c^α containing annotated word α and model c was divided by the total number of frames n_c during which the model c was most active (Eq. 7). This reveals which words were actually spoken during the activations of the model. Then the entropy was computed according to:

$$H(c) = -\sum_{\alpha=1}^{R} \frac{n_c^\alpha}{n_c} \log_R \frac{n_c^\alpha}{n_c} = -\sum_{\alpha=1}^{R} P_c(\alpha) \log_R P_c(\alpha) \qquad (7)$$

Entropy H receives a value of 0 for a fully selective model (the model reacts only to one specific word) and 1 for a totally unselective model (reacts equally to all words). R in the equation denotes the total number of different words in the reference annotation. Selectivity was computed separately for each of the models $\boldsymbol{c}$, and then the overall mean selectivity called *model entropy* H_C was computed by weighting the model entropies with the frequency f_c of occurrences of the corresponding model c in the test set:

$$H_{\mathrm{C}} = \sum_{c=1}^{N_{\mathrm{C}}} f_c H(c) / \sum_{c=1}^{N_{\mathrm{C}}} f_c \tag{8}$$

In addition, the so-called *annotation entropy* H_{A} was measured. H_{A} indicates how many alternative models have been learned (on average) for each annotated word α. H_{A} is obtained by first computing the probability distribution $P_\alpha(c)$ for each annotated word α, i.e., counting the number of frames n_α^c in which model c was active with annotated word α, and dividing it by the total number of frames n_α annotated as α. Then the entropy of these distributions is computed (Eq. 9) and the mean entropy is computed across all $\boldsymbol{\alpha}$ to obtain H_{A} (Eq. 10).

$$H(\alpha) = -\sum_{c=1}^{N_{\mathrm{C}}} \frac{n_\alpha^c}{n_\alpha} \log_R \frac{n_\alpha^c}{n_\alpha} = -\sum_{c=1}^{N_{\mathrm{C}}} P_\alpha(c) \log_{N_{\mathrm{C}}} P_\alpha(c) \tag{9}$$

$$H_{\mathrm{A}} = \frac{1}{R} \sum_{\alpha=1}^{R} H(\alpha) \tag{10}$$

If H_{A} receives a value of zero, each annotated word is represented by a single model. The more there are alternative models for words, the closer H_{A} is to the value of one. Ideally, only one model would exist for each word, capturing all varying realizations of the word. In practice this is rarely the case due to the large variability in word realizations. Finally, in order to have one descriptive measure of the modeling quality, the harmonic mean of model selectivity and model diversity, called the Q-value, can be computed by:

$$Q = 1 - \frac{2 H_{\mathrm{C}} H_{\mathrm{A}}}{H_{\mathrm{C}} + H_{\mathrm{A}}} \tag{11}$$

The Q-value integrates the information from H_{C} and H_{A}, yielding a value of $Q = 1$ only when all words in the annotation have only one model each, and that model reacts only to that specific word (ideal situation). Otherwise Q will obtain values between zero and one, where $Q =$

0 means that there is an unspecified number of word models that are activated totally randomly and independently of the acoustic input. Performance measures have been summarized in table 1.

Table 1: A summary of the performance measures used in the experiments.

Measure	Explanation
Ins	Number of additional detected segment boundaries in comparison to the reference.
σ	Mean deviation of segment boundaries from the reference (ms).
H_C	Model entropy. Indicates how selective the learned models are toward specific words. $H_C \in [0,1]$, where 0 = most selective, 1 = least selective.
H_A	Annotation entropy. Indicates the average amount of alternative models that exist for each annotated word. $H_A \in [0,1]$, where 0 = one model per word and > 0 several models per word.
Q	Overall model quality. One minus the harmonic mean of H_C and H_A. $Q \in [0,1]$, where 0 = worst possible performance and 1 = ideal performance (one fully selective model for each one word).
C	Lexical coverage, i.e., the number of detected words divided by the total number of annotated words. $C \in [0,1]$.

4. Experiments

The experiments were performed using both single talker and four talkers as training and testing material, and the results will be presented in this order. In addition, the effects of different parameters will be briefly discussed at the end of this section.

4.1 Single talker

Speech from one male talker (talker-03) was used in the first experiment to simulate speech input from one caregiver. Transition probabilities were modeled at lags $\boldsymbol{k} = \{1, 2, \ldots, 12\}$ (10-120 ms). Once the training data were preprocessed as described in section 3.1 into a long sequence of VQ labels using a codebook of $N_A = 128$ labels, the entire training signal of approximately 90 minutes was used as an input to the system. The learning procedure was performed for three different recognition thresholds t_r according to section 3.2.1. The first experiment produced a total of 30 models using $t_r = 0.033$, $L_r = 600$ ms, and $L_s = 200$ ms. The second experiment with the same windowing parameters and $t_r = 0.043$ produced a total of 113

models, and the third experiment resulted in 264 models using $t_r = 0.052$. These experiments will be referred to as the *low*, *middle*, and *high threshold* experiments.

After training, the test set of 344 novel utterances was used as an input to the recognition process (Eq. 5) and the obtained activation curves were smoothed according to Eq. (6). Figure 1 shows an example recognition of an utterance "*Do you see the square toy?*". On the top panel of the figure, a standard spectrogram of the utterance can be seen. In the middle, activation curves of all models are plotted against the same time scale as the spectrogram. In the bottom, only the winning models are retained for each moment of time, causing segmentation of the input to well-defined segments of model activity. The reference segmentation is also shown in the bottom panel as dashed lines. In this case, the segmentation has grouped "*doyouseethe*" into one long segment, as it is spoken very quickly and recurs several times in the training data. Words "*square*" and "*toy*" and silence (#h) are correctly segmented within approximately 10 ms of the reference boundaries.

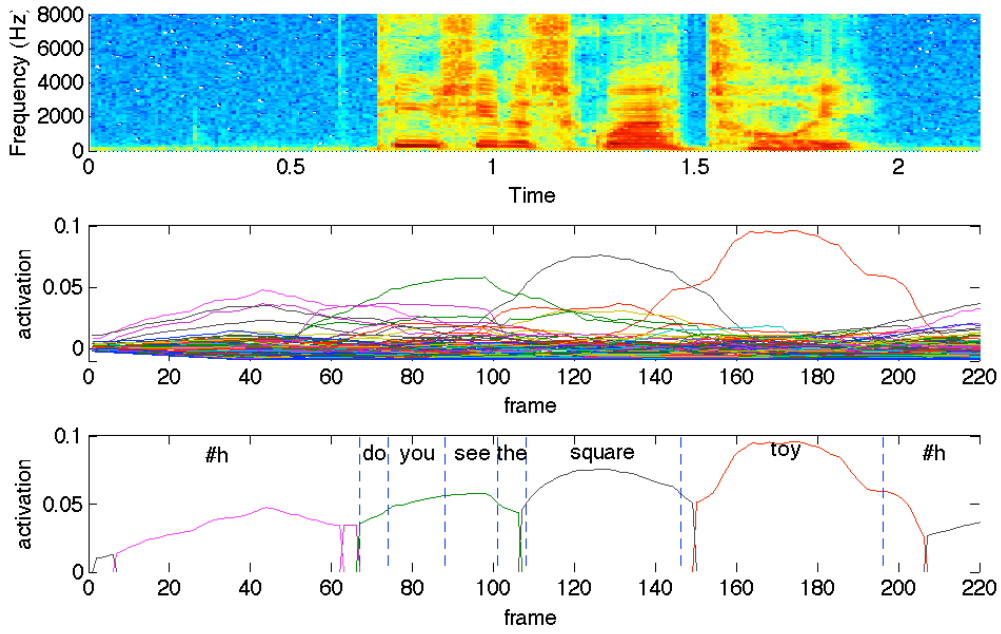


Fig. 1. A recognition example for the utterance “Do you see the square toy?”. Spectrogram of the utterance is shown at top and activation of all models are shown in the middle. At the bottom, only the winning model for each moment in time is chosen, leading to a segmentation of the input. Annotated segment boundaries (the references) are indicated by dashed lines.

Figure 2 shows the temporal measures for segmentation accuracy; the left panel shows the evolution of segmentation accuracy for the three different threshold conditions as a function of time trained while the right panel shows the corresponding insertions. The results show that the segmentation accuracy increases as more speech is perceived. However, at first this comes at the cost of introducing more segment boundaries, seen as an increase in insertion rate (the number of excess word boundaries per annotated word). However, the number of insertions does not directly correlate with reduced segment deviation, although there is a clear tendency to oversegment parts of speech that significantly overlap between different words or are otherwise ambiguous (e.g., endings of words or transitions between words). For all thresholds, the insertion rate stabilizes after 50 minutes although the segmentation accuracy keeps increasing (low and high) or stays relatively stable (middle).

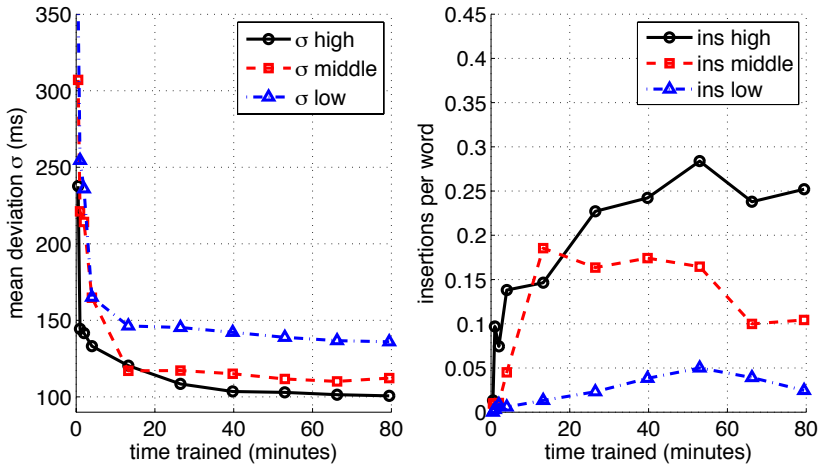


Fig. 2. Mean segment boundary deviation from reference (left) and number of insertions per word (right) as a function time trained. Three different threshold conditions are shown.

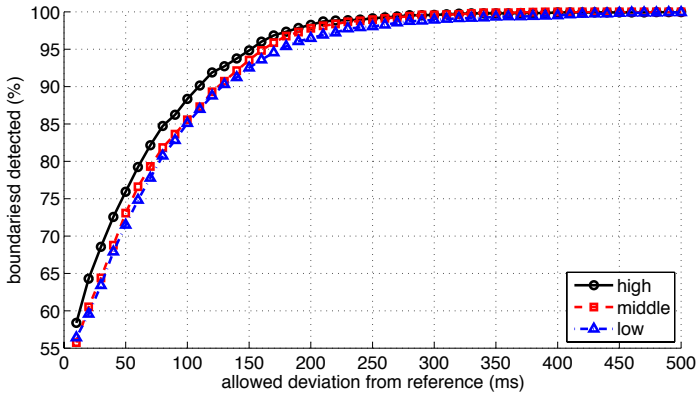


Fig. 3. The number of detected reference word boundaries (%) as a function of maximum distance that is allowed between the reference boundary and the boundary discovered by the algorithm.

Figure 3 shows the proportion of the reference boundaries that are correctly detected. Approximately 70% and 85% of boundaries are detected when ±50 ms and ±100 ms deviations are allowed (respectively) between the reference and the algorithm's output. Although it is difficult to define precisely how much deviation should be allowed for a word to be still correctly segmented, the result indicates that the majority of the word boundaries are detected within an error margin that has the same temporal scale as plosive bursts and are much shorter than the average length of vowels.

To further verify that the segmentation is still reasonable at higher insertion rates, a random segmentation was performed for the data by randomizing the temporal locations of the algorithm segment boundaries after full training in the high threshold condition (0.25 insertions per word) and computing the mean deviation from the reference. The mean deviation from the reference using the original algorithm segmentation was 100 ms (Figure 2, left), whereas the randomized case had a deviation of 159 ms. This concludes that despite the increased number of segment boundaries towards the end of training (see Figure 2, right), the algorithm performs well above chance level accuracy in word segmentation.

In order to better understand the segmentation output, the analysis has to focus on the content of the models. Figures 4 and 5 illustrate the overall entropies H_C and H_A for different recognition thresholds. Figure 6 shows the overall quality measure Q of the learning process as a function of training time. As can be seen, the entropies drop very rapidly in the beginning, and this is also the time period in which most of the models are created. A majority of the models are already in place after 10 minutes of speech and only a small number of new models are formed later. Newly created models already exhibit coarse selectivity towards acoustically similar events, and their selectivity gradually increases as more patterns are recognized and used to update the models. As the amount of training time increases, the overall selectivity of the existing models increases monotonically. As can be seen from the results, the high threshold condition leads to the most selective models in terms of both H_C and H_A. After training over the entire training signal, the overall quality of the models for the three conditions are $Q_{low} = 0.66$, $Q_{middle} = 0.73$, and $Q_{high} = 0.75$. When the high-threshold results are compared to the segmentation accuracy (Figure 2), one can observe that the number of insertions gradually increases up to the 50 minute point although the number of models per word (H_A in Figure 3) stays relatively stable after 30 minutes, suggesting that there is no significant change in the way that words are represented in the discovered models. The model selectivity H_C simply increases over time (Figure 4).

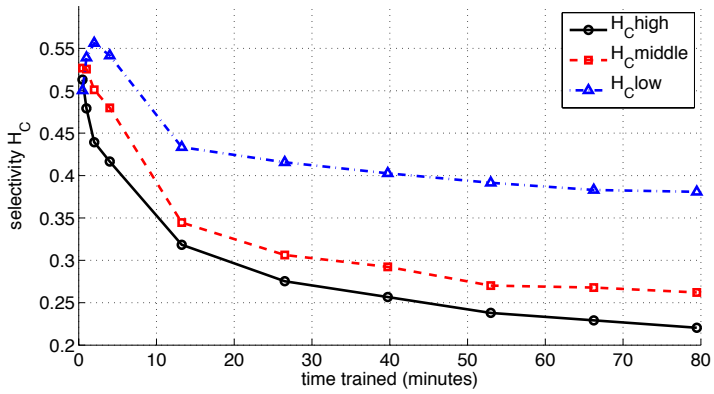


Fig. 4. Model entropy H_C as a function of time trained. Results for three different recognition thresholds t_r are shown. Lower entropy indicates higher model selectivity.

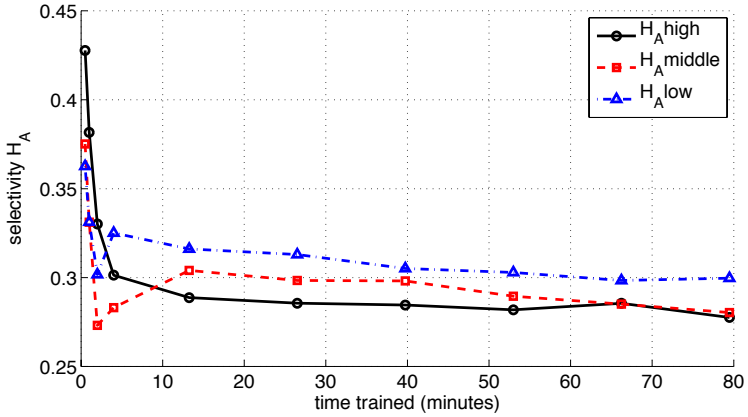


Fig. 5 Annotation entropy H_A as a function of time trained. Results for three different recognition thresholds t_r are shown. Lower entropy indicates less models for each annotated word.

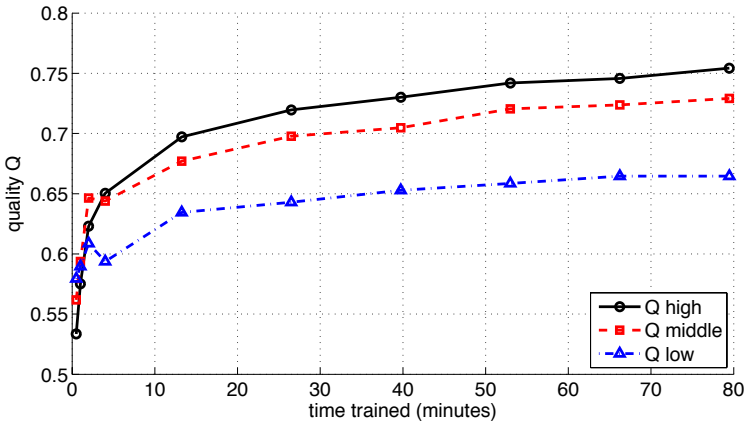


Fig. 6. Overall model quality Q as a function of time trained. Results for three different recognition thresholds t_r are shown.

Figure 7 shows a surface plot for the model selectivity H_C (Eq. 7) of individual models as a function of training time. Only 35 initially created models are shown in order to maintain the readability of the figure. As expected, the selectivity of the majority of the models improves over time (entropy decreases), although for all models the improvement is not necessarily monotonic. There are also a handful of models that actually become less selective over time. This seems to suggest that the models are not modeling only one word, but are actually specializing to a number of words (cf. the "*doyouseethe*" model in Figure 1).

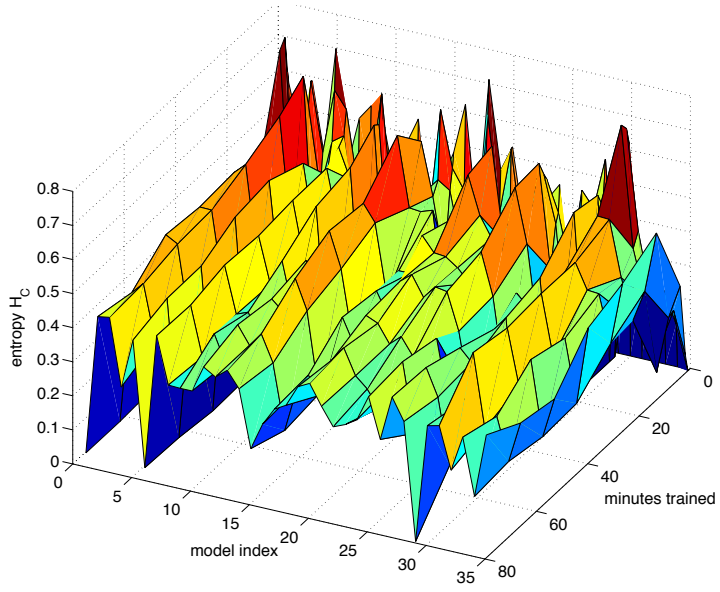


Fig. 7. *Surface-plot of model entropies H_C for the first 35 models as a function of training time for the middle threshold case. The lower the entropy, the more selective the corresponding model is towards a limited set of words.*

Lexical coverage is shown in Figure 8. The high threshold condition again produces the best results, finding approximately 85 patterns (potential words) for each 100 words in the annotation. The lexical coverage and segmentation accuracy in the low threshold condition, on the other hand, seem to indicate that too few models are learned, causing the same models to

represent several subsequent words in the speech data. In general, lexical coverage results support the earlier findings that most of the word models are already in place after ten minutes of training. However, the average model selectivity keeps increasing as more exemplars of the words occur in the data (Figures 4 and 5).

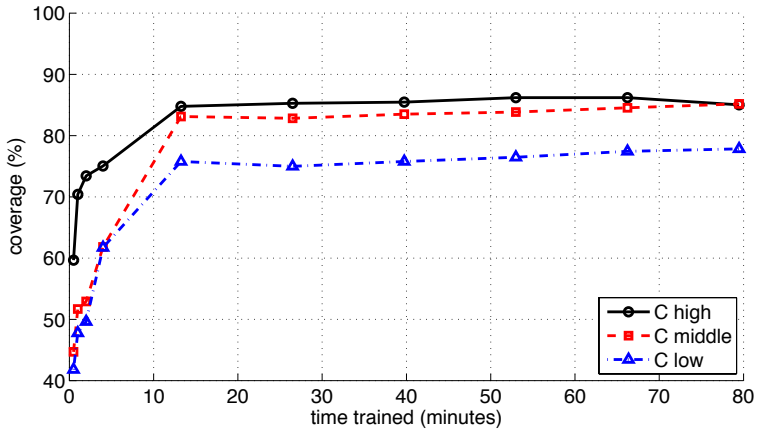


Fig. 8. Lexical coverage *C* as a function of time trained. Three different threshold conditions are shown.

Table 2 shows the contents of the most selective models in the high threshold condition. Models that are primarily reacting to silence are not shown; a full list of models and their contents are listed in Appendix C. As can be seen from the table, the selectivity of some models is relatively high. For example, one model has reacted to the word "*telephone*" fourteen times with high purity (model activation is located 95% inside the annotated "telephone" words). In addition, there are words like "*banana*", "*daddy*" and "*animal*" that include large portions of silence due to their frequent occurrence in an utterance-initial or utterance-final position. Since silence does not cause confusions between words, this simply shows up as a disagreement between the annotation and the models regarding the boundaries of utterances (e.g., refer to the

ending of model activation in Figure 1). Therefore, the overall selectivity of the models with silence can be considered to be well above 90% (e.g., banana 0.76 + #h 0.22 = 0.98).

Table 2. Contents of the most selective models in the high threshold condition. Each row describes the contents of a single model. The three most covered words are shown per model. N denotes the number of occurrences in the test set and p denotes the proportion of activation corresponding to the given word class (0-1). #h denotes silence. Primarily silence models are not shown (see Appendix C for a full list).

i	*N*	H_c	α_1	*p*	α_2	*p*	α_3	*p*
1	2	0.00	frog	1.00				
2	2	0.00	baby	1.00				
3	3	0.02	red	0.98	the	0.02		
4	2	0.02	airplane	0.98	#h	0.02		
5	2	0.05	bird	0.94	and	0.06		
6	14	0.06	telephone	0.95	and	0.02	happy	0.01
7	5	0.09	airplane	0.91	and	0.05	a	0.03
8	4	0.11	sad	0.89	the	0.05	a	0.03
9	5	0.12	said	0.87	small	0.06	square	0.04
10	6	0.12	dirty	0.87	the	0.07	is	0.02
11	6	0.12	bird	0.87	#h	0.04	a	0.04
12	4	0.13	edible	0.82	the	0.15	sees	0.03
13	6	0.13	duck	0.87	red	0.04	and	0.04
14	7	0.14	happy	0.84	the	0.11	gives	0.02
15	6	0.14	cookie	0.85	square	0.07	and	0.03
16	9	0.14	airplane	0.78	#h	0.20	happy	0.01
17	12	0.15	banana	0.76	#h	0.22	and	0.02
18	9	0.15	bottle	0.85	#h	0.06	and	0.04
19	8	0.15	round	0.83	gives	0.09	a	0.07
20	3	0.16	telephone	0.79	big	0.12	red	0.05
21	2	0.17	telephone	0.72	dirty	0.21	big	0.07
22	3	0.18	apple	0.76	and	0.13	happy	0.07
23	3	0.18	daddy	0.72	#h	0.19	gives	0.07
24	12	0.18	animal	0.71	#h	0.24	an	0.03

The least selective models of the high threshold condition are listed in Table 3. It is difficult to say anything comprehensive about why these models fail to represent any single word, but possible reasons include the fact that two or more short words end up easily inside the same analysis window during learning, leading to a common representation for all of the words.

Such examples include model 13, "*[do] [you] [like]*", model 9 "*[where] [is]*", and model 4 "*[have you]*" or "*[(do) you have]*". In addition, there are some words that share acoustic similarities, e.g., "*ball*" and "*small*" (model 8), "*cow*" and "*cat*" (model 11), and "*red*" and "*round*" (model 6). It is also simply possible that the analysis window has detected a novel segment of speech that spans partially across two longer words that happen to occur subsequently (e.g., …*cle*{*an fro*}*g* in model 15, window denoted with {}). As new occurrences of these words (*cleans* and *frogs*) occur in isolation, this model may still achieve sufficiently high activation to become updated, now with the full temporal span of the word. In this manner, the model becomes gradually more and more selective for full forms of the both words. Since the current implementation does not perform pruning, splitting, or merging of the representations, these errors do not become corrected.

Table 3. Contents of the least selective models in the high threshold condition. The three most frequent words are shown per model. *N* denotes the number of occurrences in the test set and *p* denotes the proportion of activation corresponding to the given word class (0-1). #h denotes silence.

i	*N*	H_C	α_1	*p*	α_2	*p*	α_3	*p*
1	29	0.50	see	0.26	you	0.21	#h	0.13
2	3	0.48	see	0.19	happy	0.15	the	0.15
3	7	0.47	animal	0.27	horse	0.14	apple	0.14
4	12	0.45	have	0.35	you	0.16	#h	0.14
5	15	0.45	man	0.34	smiling	0.22	round	0.1
6	10	0.44	red	0.31	round	0.19	is	0.13
7	3	0.44	have	0.29	the	0.18	you	0.16
8	22	0.41	ball	0.45	porsche	0.16	small	0.15
9	25	0.39	where	0.43	#h	0.15	is	0.13
10	7	0.38	crying	0.33	round	0.29	happy	0.13
11	8	0.38	cow	0.35	cat	0.22	#h	0.19
12	4	0.38	looks	0.4	big	0.18	at	0.15
13	16	0.37	like	0.35	you	0.21	do	0.18
14	6	0.37	happy	0.43	gives	0.21	has	0.11
15	10	0.37	clean	0.4	frog	0.26	truck	0.15

Table 4. All annotated words occurring in the test set. Word specific annotation entropies H_A, three most activated models *i,* and the corresponding proportions *p* (0-1) of total word durations are shown for the high threshold condition.

		1st		2nd		3rd	
α	H_A	*i*	*p*	*i*	*p*	*i*	*p*
#h	0.53	2	0.16	15	0.15	92	0.15
a	0.76	133	0.04	109	0.04	19	0.03
airplane	0.26	121	0.44	116	0.29	31	0.12
an	0.44	36	0.27	7	0.11	40	0.10
and	0.61	20	0.08	109	0.07	52	0.06
animal	0.28	7	0.51	83	0.19	30	0.11
apple	0.33	83	0.43	167	0.17	69	0.15
at	0.31	43	0.45	109	0.18	104	0.13
baby	0.41	200	0.32	80	0.16	172	0.10
ball	0.15	20	0.66	4	0.30	135	0.01
banana	0.23	151	0.52	53	0.22	106	0.17
big	0.57	213	0.09	44	0.07	102	0.07
bird	0.30	107	0.32	67	0.28	55	0.18
blue	0.35	136	0.41	72	0.16	53	0.15
bottle	0.25	192	0.50	152	0.21	83	0.18
car	0.20	130	0.53	23	0.32	67	0.09
cat	0.23	32	0.41	103	0.35	102	0.17
clean	0.29	23	0.52	120	0.20	77	0.11
cookie	0.19	28	0.50	84	0.41	184	0.04
cow	0.26	32	0.38	103	0.33	102	0.17
crying	0.10	47	0.75	93	0.25	1	0.00
daddy	0.40	64	0.25	181	0.17	188	0.10
dirty	0.47	39	0.26	42	0.20	170	0.05
do	0.37	57	0.24	40	0.22	106	0.16
dog	0.25	170	0.66	88	0.06	118	0.06
doll	0.23	170	0.67	34	0.10	8	0.06
duck	0.19	63	0.46	88	0.44	126	0.06
eagle	0.22	36	0.49	83	0.35	165	0.06
edible	0.26	59	0.47	205	0.28	195	0.13
fish	0.06	48	0.94	16	0.03	123	0.02
frog	0.31	100	0.42	201	0.19	38	0.12
gives	0.44	129	0.18	147	0.18	95	0.17
happy	0.44	25	0.33	10	0.15	99	0.12
has	0.41	188	0.27	61	0.13	218	0.12
have	0.30	125	0.43	76	0.24	32	0.09
he	0.46	80	0.14	36	0.13	57	0.11

		1st		2nd		3rd	
α	H_A	*i*	*p*	*i*	*p*	*i*	*p*
here	0.24	91	0.40	117	0.31	54	0.20
horse	0.33	18	0.37	4	0.22	180	0.12
I	0.12	60	0.69	108	0.31	169	0.01
is	0.48	22	0.20	91	0.12	17	0.10
like	0.21	40	0.45	133	0.39	124	0.10
likes	0.35	74	0.23	13	0.21	14	0.19
lion	0.30	52	0.41	158	0.25	45	0.11
looks	0.28	43	0.52	50	0.13	104	0.11
man	0.32	110	0.38	19	0.29	50	0.10
mean	0.03	108	0.97	52	0.03	35	0.00
mummy	0.26	6	0.47	61	0.34	222	0.05
no	0.09	60	0.84	70	0.13	108	0.03
porsche	0.26	4	0.43	20	0.27	2	0.14
red	0.44	101	0.20	150	0.19	22	0.14
robin	0.29	75	0.28	152	0.26	132	0.24
round	0.43	137	0.27	134	0.18	22	0.09
sad	0.34	49	0.36	96	0.19	89	0.16
said	0.21	169	0.58	60	0.29	109	0.04
see	0.19	57	0.65	106	0.24	159	0.05
sees	0.44	26	0.33	123	0.13	100	0.10
she	0.44	157	0.12	66	0.12	86	0.11
small	0.26	12	0.45	16	0.25	20	0.17
smiling	0.16	35	0.66	19	0.25	164	0.07
square	0.36	122	0.29	79	0.26	11	0.14
takes	0.21	9	0.71	228	0.09	110	0.06
telephone	0.23	21	0.66	211	0.12	149	0.07
the	0.74	25	0.05	22	0.05	57	0.05
there	0.43	54	0.21	141	0.18	8	0.10
toy	0.22	85	0.52	34	0.34	69	0.05
tree	0.22	135	0.46	16	0.33	109	0.16
truck	0.22	109	0.68	120	0.09	213	0.08
where	0.14	22	0.59	17	0.39	8	0.01
woman	0.37	94	0.41	119	0.13	147	0.08
yellow	0.40	162	0.24	180	0.24	140	0.10
you	0.36	57	0.33	40	0.18	106	0.16

Table 4 lists all words occurring in the test set, the corresponding annotation entropies H_A, and three models *i* that are most active during the word, and the relative temporal coverage *p* of these models during the high threshold condition. The table shows that several models exist for the majority of the annotated words, while the words "*mean*", "*fish*", and "*no*" are exceptions to the rule. Words such as "*I*", "*ball*", "*car*", "*duck*", "*eagle*", "*see*" and "*toy*" are mainly represented by two different models since the two models cover nearly 100% of the word occurrences in the test set. On the other hand, very short function words and verbs such as "*a*", "*an*", "*the*", "*is*", "*do*" and the silence "#h" activate a large variety of models. By combining the information from Table 3, it can be concluded that these short and often sloppily pronounced words hardly obtain their own representations in the system, but are included in larger proto-lexical constructs such as "*a ball*", "*an apple*", "*doyousee*" (Figure 1) or "doyoulikethe" (Appendix D).

Finally, Figure 9 shows the distributions of the word durations, both for the learned words (red) and annotated words (blue) (silence segments are excluded). As can be seen, there is an evident tendency for the algorithm to produce somewhat shorter segments than have been annotated, although a large number of long (> 150 ms) segments can also be found. Both distributions are clearly bimodal, showing that the algorithm is segmenting both mono- and multisyllabic words. The insertion rates for monosyllabic words were 8.2% for the middle threshold case and 20% for the high threshold case, whereas multisyllabic words had respective insertion rates of 11.7% and 25.2%. In general, it seems that despite a fixed analysis window size of 600 ms, the learned models vary notably in their duration and the distribution of durations approximately follows the distribution derived from the annotation. A relatively low increase in

insertion rate from mono- to multisyllabic words also supports the finding that words are mainly modeled in their entirety instead of being chunked into syllable-like units.

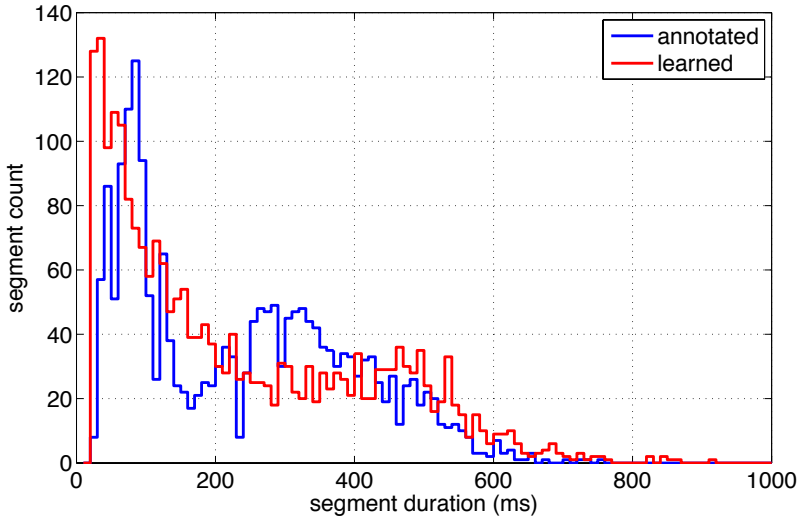


Fig. 9. Distributions of segment lengths for learned segments (red) and annotated segments (blue).

Manual perceptual evaluation supports the findings from automatic evaluation. When the detected speech segments are extracted, then categorized according to models, and finally listened to, many of the models exhibit fairly accurate word segmentation in terms of subjective perceptual judgment (see also some spectrogram examples in Appendix D). As can be expected based on Table 2, some of the models are very pure and only rarely contain extraneous signal contents in addition to one specific word, whereas some of the models are selective and accurate simultaneously for two different words, or combinations of two or more short words that occur often consecutively.

Overall, the results clearly show that in the case of only one talker and with the small vocabulary used in the experiments, learning of coarse word-like models is rapidly achieved. Model selectivity and segmentation accuracy are already greatly above chance level after a few

word tokens, and only a slight gradual improvement can be perceived as more speech is introduced.

It is also evident from Tables 3 and 4 that the number of learned models is notably higher than the true number of words in the vocabulary even though the selectivity of the models is not ideal. However, this is not surprising considering the complexity of real speech. Context dependent and normal intra-speaker variation can cause significant acoustical changes to the signal, yielding different representations in the used discrete acoustic space. It is also possible that two or more word models develop towards very similar representations as more training data are introduced, and they both become activated once their characteristic word is spoken. Since the current implementation has no special mechanism for pruning or merging of models, similar models will remain as parallel alternatives for the same word. In the future it would be worthwhile to study grouping and segmenting of models in terms of isomorphic properties of their activation curves.

4.2 Multiple talkers

The same experiment as described in section 4.1 was also performed for four talkers (two males, two females). There were two separate conditions: a random case, where all utterances from all four talkers were mixed in a random order, and in addition, the experiment was repeated in a talker-blocked order (male, female, male, female), in which the first 2144 utterances were from talker-01, the next 2144 from talker-02 and so on. The test set always contained 1376 randomly chosen novel utterances from all four talkers, although the results are also reported separately for each talker in the test set. In the following figures, the results from the single-speaker experiment are also plotted for reference purposes using a black dashed line.

A VQ codebook of size $N_A = 128$ labels was trained using randomly chosen speech material from the training set. By using the high threshold level of the single talker case ($t_r = 0.052$), the full training signal of approximately 335 minutes (over 5.5 hours) in length was used as input to the learning algorithm. This produced a total of 287 and 304 models for the random and blocked cases, respectively. These numbers are slightly higher than the 264 models in the single talker case, although the alphabet still consists of the same 80 words.

Figure 10 shows the segmentation accuracies in terms of mean deviation from reference and Figure 11 shows the number of insertions per word for each talker. The first observation is that a randomized talker order leads to faster convergence of the mean deviation for all talkers, whereas a blocked ordering requires full training of the two first talkers to achieve an accuracy comparable to the single-speaker condition. Interestingly, there are also small bumps in the mean boundary deviation after the introduction of a new speaker. When the insertions are studied, there are notable differences between blocked and randomized orders: with blocked order the insertion rates for talker 3 and 4 data increase gradually to relatively high levels during the training of the previous talkers, but drop down quickly when speech from the corresponding talkers are introduced to the system. Both the deviation bumps and the quickly dropping insertion rates are caused by a number of new models that are learned at talker change points in order to address the mismatch between existing models and the new data. These new models have initially poor selectivity, causing inaccuracies in the segmentation. However, as already noted in the single-speaker experiments, approximately 10 minutes of data from a new talker are sufficient to improve the model quality to a large degree, stabilizing the segmentation procedure. The results also suggest that data from the two first talkers generalize poorly to the remaining two caregivers, causing words to be split into smaller sub-segments and therefore increasing the

number of insertions. In general, the mean segmentation accuracies fall slightly behind the accuracy achieved within the single talker experiment, but the differences are not large. This is also verified in Figure 12 that reveals the mean number of correctly detected reference boundaries over all four talkers.

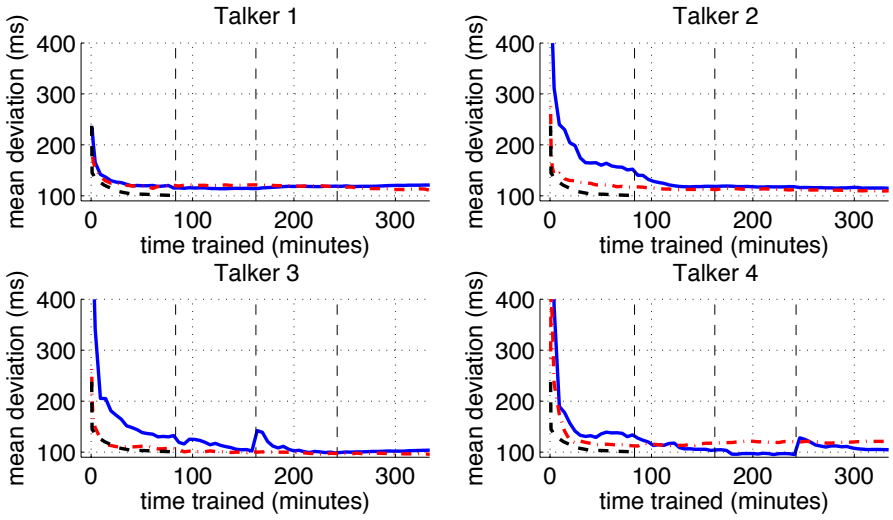


Fig. 10. Mean segment boundary deviation from reference. Results are shown separately for each talker. The blue solid line and red dotted line correspond to blocked and randomized conditions (respectively). The result from the single speaker experiment is shown as a reference using a black dashed line. Talker change locations for blocked ordering are shown using dashed vertical lines.

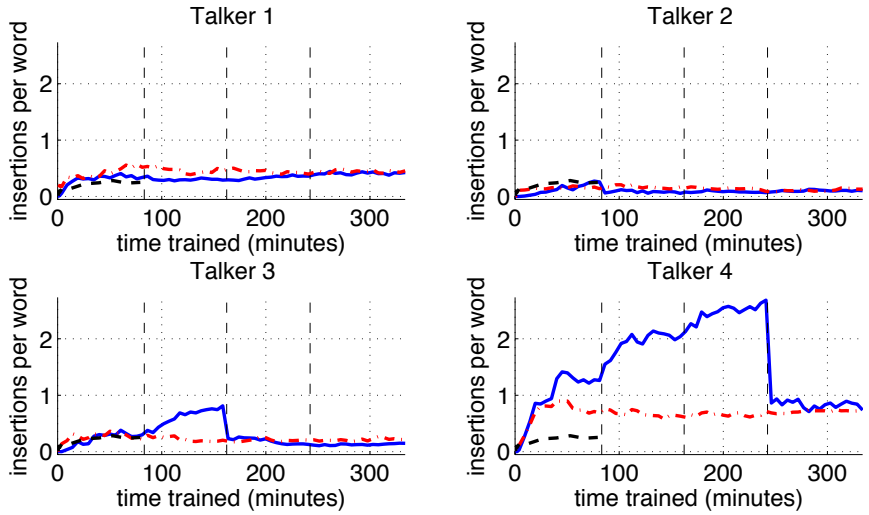


Fig. 11. The mean number of insertions per annotated word. Results are shown separately for each talker. The blue solid line and red dotted line correspond to blocked and randomized conditions (respectively). The result from the single speaker experiment is shown as a reference using a black dashed line. Talker change locations for blocked ordering are shown with vertical dashed lines.

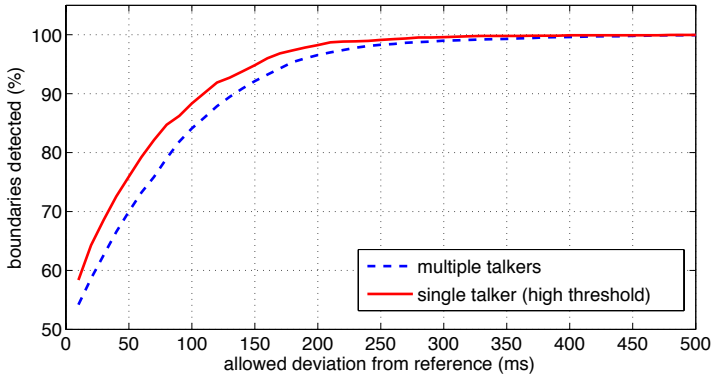


Fig. 12. The number of detected reference word boundaries (%) as a function of maximum distance that is allowed between the reference boundary and the boundary discovered by the algorithm.

As for the model contents, Figure 13 shows the mean model selectivity in terms of model entropy H_C. The talker specific differences in speech are readily seen. Since the test set contains speech from all four talkers, the models learned in the talker-blocked order are not very good at classifying words from new talkers if only one or two talkers are seen. During exposure to the first talker, there is hardly any improvement in the model qualities for speech from the other talkers. When speech from the second talker is used to train the system at the first block change point, the model selectivity starts to improve quickly for the second talker. There is also a much better generalization from talker 2 to talker 3 than there is from talker 1, and this is seen as a steeper slope in the entropy curve in the lower left panel during the second block. Still, a significant speedup in the evolution of model quality can be perceived for talker 3 when speech material from talker 3 is also used to train the system during the third block. Interestingly enough, the learning speed for talker 1 is not significantly hindered during the randomized

condition, although learning is naturally much faster for all the remaining three talkers as they are also included in the training data immediately from the start.

The differences in the verbal characteristics of the four talkers are also reflected in Figure 14 that describes how many parallel models exist for each annotated word. Since the selectivity of the models is poor for novel talkers, a very high number of models are reacting to each annotated word (high annotation entropy H_A) as long as there is no speech from the corresponding talker in the training data. As soon as matching data become available, the number of parallel models drops quickly as models can adapt themselves to the talker specific patterns. In terms of annotation entropy, the generalization from the first three talkers to the last talker is much worse than in terms of model selectivity (cf. Figure 13). This suggests that many of the words spoken by talker 4 can be recognized relatively well using the models learned on the basis of talkers 1-3, but there are several – possibly speaker-specific – models from which one or another word is activated depending on the given realization of the word. Only the matching training data from talker 4 are able to limit the number of parallel models to a much smaller set, either by creating new models for some of the words, but more often updating a subset of the existing ones to account for the spectrotemporal characteristics of the new talker. The generalization problems from one speaker to another are also reflected in the overall model quality measure Q in Figure 15.

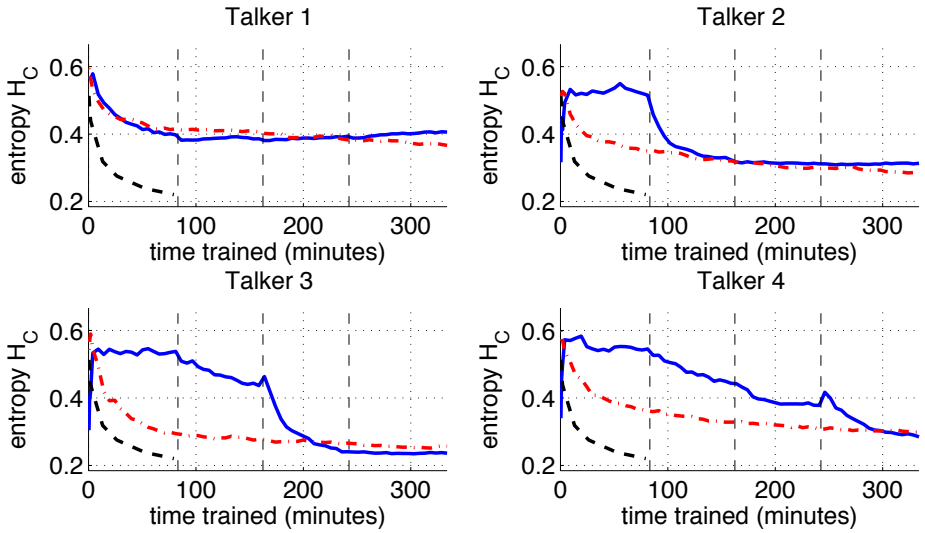


Fig. 13. Model entropy H_C as a function of time trained. Results for randomized talker order (red dotted line) and talker blocked ordering (blue solid line) are shown separately. Talker change locations for blocked ordering are shown using dashed vertical lines. In addition, the H_C value from the single talker experiment is shown as a reference using a black dashed line

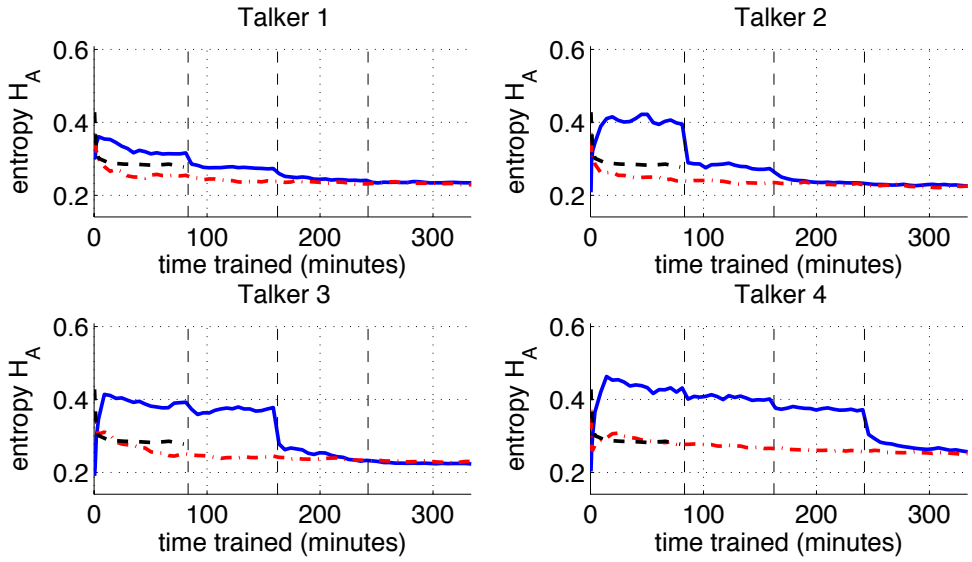


Fig. 14. Annotation entropy H_A as a function of time trained. Results for randomized talker order (red dotted line) and talker blocked ordering (blue solid line) are shown separately. Talker change locations for blocked ordering are shown using dashed vertical lines. In addition, the H_A value from the single talker experiment is shown as a reference using a black dashed line

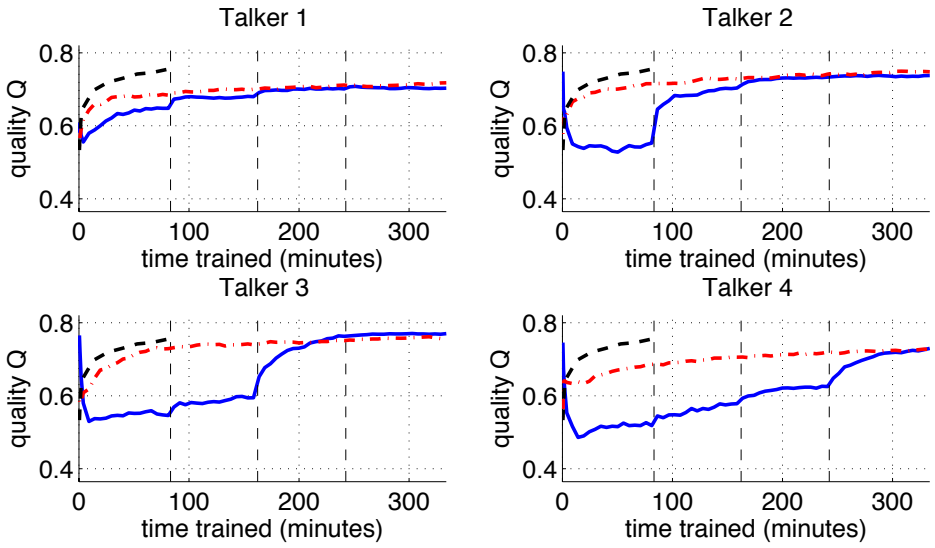


Fig. 15. Overall model quality *Q* as a function of time trained. Results for randomized talker order (red dotted line) and talker blocked ordering (blue solid line) are shown separately. Talker change locations for blocked ordering are shown using dashed vertical lines. In addition, the *Q* value from the single talker experiment is shown as a reference using a black dashed line.

In order to get an overview of the process instead of talker specific performance, Figure 16 indicates the model entropy H_C, annotation entropy H_A, and the Q measure evaluated over the entire test set of four talkers. Note that the entropy values are not averages of talker specific values because the model activations are not now studied in isolation from the speech by other talkers (see Equations 7-10). This leads to higher entropies due to acoustically overlapping but linguistically distinct patterns between talkers (e.g., the same model may represent different words or parts of words for different talkers). In other words, the internal representations are less organized in terms of linguistic identities of the patterns if talker identity is not available to the system during the recognition of a pattern. This means that multiple-speaker performance is much worse than that of the single-speaker experiment (black dashed lines) when the correspondence between mappings from acoustic patterns to annotated words is measured in the

context of speaker-independent models. Although the quality of the models keeps increasing to the end of training, the overall Q measure obtains only a value of Q = 0.62 in the randomized talker order in contrast to the Q = 0.75 obtained in the single-speaker conditions during 90 minutes of training data.

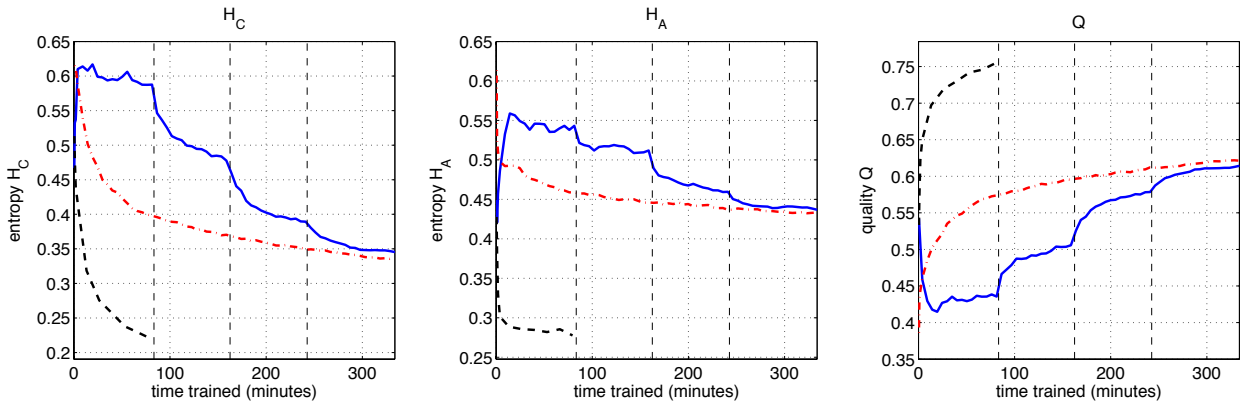


Fig. 16. Model entropy H_C (left), annotation entropy H_A (middle), and overall quality Q (right) evaluated over the entire multi-talker test set. Results for randomized talker order (red dotted line) and talker blocked ordering (blue solid line) are shown separately. Talker change locations for blocked ordering are shown with vertical lines. Single-speaker result is shown for comparison with the dashed black line.

Table 5 shows the contents of the eight most selective models (top section). Although the selectivity of these models is good, one immediately notices that the occurrence count of these models is relatively low (note that the total number of words in the material is four times than that of the single talker condition). The bottom section of Table 5 shows a number of models that have a higher number of occurrences. The selectivity of these models is already notably worse than for the most selective models, but still comparable to the single talker condition. Table 6 shows the distribution of model activities for each annotated word. The increased number of parallel models per each keyword in the multi-talker experiment can be seen in the word specific

model distributions since nearly all words have a higher annotation entropy H_A than in the single-speaker case.

Finally, Figure 17 shows the lexical coverage of the multitalker experiments. Lexical coverage follows the same trend as the other measures, randomized talker order performing better with less data, as speech from all talkers in the test set is taken into account already at the beginning of the training. After full training, the performance in the blocked order experiment reaches similar levels to those measured when randomized order was applied.

Table 5: Contents of models in the multiple talker experiment. Eight most selective models (in terms of entropy) are shown. For each model, *p* denotes the proportion (0-1) of each word α. Additionally, a number of relatively selective models with a high occurrence count *N* are listed.

i	*N*	H_C	α_1	*p*	α_2	*p*	α_3	*p*
1	4	0.04	truck	0.96	robin	0.04		
2	2	0.06	here	0.93	#h	0.07		
3	3	0.08	telephone	0.9	a	0.07	and	0.02
4	8	0.09	cookie	0.87	#h	0.13		
5	8	0.10	dirty	0.91	a	0.03	bottle	0.03
6	2	0.10	banana	0.84	and	0.16		
7	10	0.12	dog	0.88	big	0.05	clean	0.03
8	2	0.12	sad	0.85	the	0.09	has	0.07
	...		...	...	...	...	...	...
9	11	0.17	gives	0.76	she	0.12	he	0.11
10	10	0.18	telephone	0.76	woman	0.12	happy	0.09
11	23	0.18	happy	0.83	yellow	0.05	mummy	0.04
12	13	0.21	telephone	0.73	sad	0.15	yellow	0.06
13	17	0.22	man	0.77	sees	0.05	a	0.05
14	14	0.23	bottle	0.76	big	0.06	square	0.06

Table 6. All annotated words occurring in the multi-talker test set. Word specific annotation entropies H_A, the three most activated models *i*, and the corresponding proportions *p* (0-1) of word durations are shown.

		1st		2nd		3rd	
α	H_A	*i*	*p*	*i*	*p*	*i*	*p*
#h	0.62	133	0.11	64	0.08	80	0.08
a	0.84	21	0.04	18	0.03	65	0.03
airplane	0.59	140	0.13	142	0.09	71	0.07
an	0.66	244	0.11	12	0.05	18	0.04
and	0.57	244	0.18	4	0.13	74	0.13
animal	0.56	87	0.13	41	0.09	125	0.08
apple	0.49	65	0.15	5	0.14	167	0.13
at	0.39	114	0.25	78	0.20	98	0.15
baby	0.51	92	0.18	149	0.16	69	0.12
ball	0.53	52	0.12	40	0.11	149	0.09
banana	0.56	136	0.17	203	0.09	112	0.09
big	0.63	145	0.12	182	0.12	156	0.07
bird	0.56	91	0.16	290	0.12	201	0.12
blue	0.58	60	0.18	177	0.12	224	0.08
bottle	0.46	77	0.17	65	0.16	270	0.16
car	0.40	68	0.28	255	0.25	202	0.10
cat	0.52	18	0.15	38	0.13	41	0.13
clean	0.47	51	0.24	70	0.13	156	0.11
cookie	0.34	185	0.29	170	0.25	59	0.21
cow	0.55	211	0.15	68	0.14	38	0.10
crying	0.33	35	0.26	45	0.23	51	0.22
daddy	0.59	149	0.18	205	0.07	31	0.07
dirty	0.56	10	0.18	176	0.11	194	0.08
do	0.40	50	0.27	2	0.20	14	0.12
dog	0.52	150	0.18	40	0.11	65	0.10
doll	0.48	52	0.26	40	0.15	82	0.09
duck	0.51	99	0.18	213	0.16	4	0.12
eagle	0.50	83	0.22	167	0.09	70	0.09
edible	0.52	216	0.21	180	0.13	84	0.08
fish	0.34	133	0.34	66	0.27	118	0.18
frog	0.37	42	0.33	161	0.20	65	0.15
gives	0.52	36	0.17	205	0.15	154	0.10
happy	0.51	15	0.19	141	0.16	208	0.15
has	0.44	28	0.21	60	0.20	39	0.17
have	0.40	16	0.21	28	0.19	50	0.16
he	0.51	59	0.26	148	0.07	21	0.06

		1st		2nd		3rd	
α	H_A	*i*	*p*	*i*	*p*	*i*	*p*
horse	0.49	34	0.16	19	0.14	91	0.10
I	0.33	151	0.32	26	0.19	93	0.18
is	0.58	21	0.10	189	0.09	24	0.09
like	0.40	126	0.29	37	0.20	94	0.12
likes	0.40	79	0.28	72	0.23	199	0.13
lion	0.54	11	0.16	224	0.10	41	0.09
looks	0.44	114	0.25	78	0.17	98	0.16
man	0.52	134	0.20	18	0.19	143	0.06
mean	0.50	26	0.20	151	0.12	33	0.11
mummy	0.57	25	0.13	106	0.11	199	0.10
no	0.37	158	0.29	151	0.21	93	0.15
porsche	0.43	67	0.23	93	0.18	27	0.16
red	0.57	132	0.11	240	0.10	223	0.08
robin	0.56	46	0.13	11	0.09	197	0.08
round	0.59	113	0.12	68	0.11	100	0.11
sad	0.44	18	0.19	16	0.18	47	0.14
said	0.40	151	0.24	93	0.15	26	0.15
see	0.27	2	0.47	69	0.29	14	0.10
sees	0.43	31	0.17	2	0.17	69	0.17
sells	0.00	1	0.00	2	0.00	3	0.00
she	0.24	17	0.71	72	0.05	148	0.04
small	0.42	27	0.25	95	0.18	65	0.15
smiling	0.45	45	0.20	51	0.17	22	0.13
square	0.46	7	0.22	30	0.16	58	0.11
takes	0.41	22	0.22	108	0.19	59	0.18
telephone	0.46	107	0.22	162	0.21	117	0.12
the	0.79	2	0.07	69	0.04	18	0.03
there	0.50	34	0.19	189	0.14	146	0.11
this	0.26	208	0.31	189	0.24	29	0.23
toy	0.45	49	0.18	183	0.17	43	0.14
tree	0.36	229	0.26	85	0.23	119	0.18
truck	0.47	85	0.25	213	0.13	62	0.09
where	0.37	30	0.22	190	0.17	109	0.14
woman	0.46	25	0.17	97	0.17	111	0.14
yellow	0.55	20	0.14	186	0.13	26	0.08
you	0.41	50	0.27	2	0.14	14	0.13

here	0.40	21	0.28	24	0.16	47	0.12

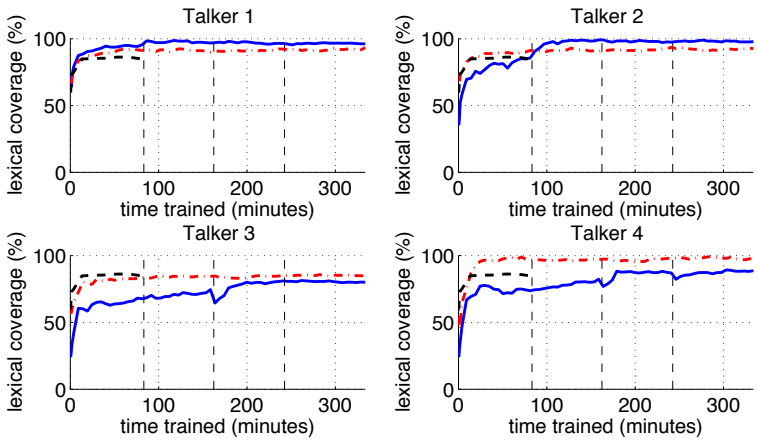


Fig. 17. Lexical coverage *C* as a function of time trained. Results for randomized talker order (red dotted line) and talker blocked ordering (blue solid line) are shown separately. Talker change locations for blocked ordering are shown using dashed vertical lines. In addition, the *C* value from the single talker experiment is shown as a reference using a black dashed line.

Overall, the results show that segmentation performance is relatively good also for the case of multiple, acoustically distinct talkers (Figures 10-12). The learned models also show coarse selectivity towards specific words, although the modeling performance is far below the single talker case even after full training. This is somewhat the expected result since the constant number of VQ-labels for acoustic events starts to correspond to a broader variety of speech sounds as more talkers are introduced. This makes the atomic representation of speech less detailed and greatly reduces the transition probability marginal between familiar and novel patterns.

In general, the results from multiple talkers are in line with what is known about acoustic differences between talkers and thereby mismatches between talker specific models – a problem that is causing large challenges in the task of automatic speech recognition (e.g., Huang, 1992).

This is also in line with the literature regarding the first stages of infant speech perception. It is known that at 9 months of age, an infant's representations of speech are overly detailed, and differences in pitch (Singh et al., 2008), sex (i.e., pitch, formant frequencies and possibly speaking style; Houston & Jusczyk, 2000), and stress (Bortfeld & Morgan, in press) in the spoken word tokens have an adverse effect on their word recognition performance. Only later in development do infants become proficient in generalizing across talkers. Regarding the gender of the talker, infants of age 10.5 months were able to perform this generalization (Houston & Jusczyk, 2000).

4.3 Effects of the parameters

A general drawback with the proposed model for word segmentation is that all the above parameters need to be defined manually and a poor selection of parameters can lead to low quality word models. The model activity threshold t_r determines the number of models that are learned for the given data, whereas the window length L_r affects the learning speed and characteristic lengths of the units that are learned. If L_r is set too low, the decision between *familiar* vs. *novel* pattern has to be made based on less data and the window never sees longer words in full. This leads to a modeling of short recurring acoustic segments that have high selectivity, and the words become fragmented into short sub-word segments. This increases the number of insertions radically if word level annotation is used as a reference point. On the other hand, a too large L_r potentially leads to a situation where several novel words that occur subsequently in the training data are incorporated into a single model. This is because both words might fit into the analysis window at the same time and therefore the newly formed model might become the dominating model for both of the words.

Other affecting factors include the codebook size and the way that activation values are smoothed temporally. For the given corpus, the overall performance was not highly affected when the codebook size was varied between 32 and 150 labels. Larger codebooks simply require lower novelty thresholds t_r since the probabilities of spectrotemporal trajectories become smaller as the number of different possible trajectories increases. At the extremes of codebook size, things naturally change. For very small codebooks the spectral resolution is not sufficient to differentiate between different words and model selectivity becomes poor. For very large codebooks ($N_A >> 256$) the resolution becomes too high to detect recurring structures, yielding totally different VQ sequences for even slightly different realizations of a word.

The issue of temporal smoothing of activation values is more complicated. During novel/old classification they do not play any role since the winning model is simply the one with the highest cumulative activation within the given time window L_r. During final recognition and segmentation, however, smoothing has a significant effect. Without smoothing, the winning word model can change very rapidly from one model to another and then back again, e.g., in the case where there are several competing words sharing the same syllabic structures. This "jumping" from one model to another causes a high amount of over-segmentation if every winning model change point is defined as a word boundary. The effect disappears when sufficient temporal filtering is applied. On the other hand, too much smoothing has an adverse effect on the temporal accuracy. In this work we used a simple moving-average (SMA) filtering technique of length 480 ms since it was found to lead to reasonable results (see also Räsänen et al., 2009a). This choice was not motivated by any existing theory, although similar smoothing mechanisms are assumed in, e.g., in TRACE model of speech perception where the phoneme and word unit activations decay over time (McClelland & Elman, 1986).

5. Discussion

The present work demonstrates that automatic word segmentation and learning of primitive ungrounded lexical items from real continuous speech is possible without pre-existing linguistic knowledge (e.g., a phonemic system) or contextual support by simply analyzing transitional probabilities between atomic acoustic events. Moreover, the current computational model demonstrates how the acquisition of protolexical word models aids, and is parallel to, the word segmentation task. This provides support to the distributional learning hypothesis (e.g., Saffran et al., 1996; Saffran, 2001) and the idea that preliminary lexical items might precede the formation of a phonemic re-organization of perception, as is suggested in the PRIMIR theory of language acquisition (Werker & Curtin, 2005).

Note that the model does not prove that the phonemic categories follow primitive lexical learning (cf. Werker & Curtin, 2005; Pisoni, 1997; Port, 2007), nor that real infants would perceive speech as a sequence of discrete elements, but simply shows that the knowledge of phonemic categories, or even segments of phones or syllables, is not necessary for rudimentary lexical learning. As discussed in the introduction, the assumption of phonemic perception before lexical learning is controversial and it is more likely that if phonemic representation exists at all, it develops with the help of some kind of proto-lexical layer that provides the necessary constraints for the development of categorical perception of sub-word structures (cf. Feldman et al., 2009). The current results are also in conflict with the work of Yang (2004), where it is explicitly claimed that transitional probability analysis alone cannot segment words from real speech, but innate linguistic constraints are necessary to determine what properties of the signal should be attended to.

The results from infant literature and also from the experiments described in this paper seem to indicate that young infants may originally have representations of words that are tied closely to acoustic details and describe the words or even combinations of often co-occurring words as a whole. The awareness that words are constitutes of smaller invariant building blocks such as phonemes is not yet in place nor required for lexical learning. Simply stated, word recognition takes place if a sufficiently close acoustic match to an existing representation occurs. The absence of a phonemic system also means that two or more similar words (e.g., minimal pairs) may become integrated into a single proto-lexical model since their similarity may be higher than the similarity of the same word spoken by two different persons (cf., e.g., Rost & McMurray, 2009). Similarly, an identical word spoken in different contexts by one speaker, or by different speakers, may lead to several parallel items in the lexicon, as long as there are no additional cues that would signify categorical similarity of the tokens. Only experience, not only with more perceived speech, but also with the functional and social contexts in which the speech takes place, can provide the necessary support for the development of lexical and sub-lexical representations that can overcome the limitations of the initial models that are based on acoustic regularities.

As for the methodological aspect, the proposed algorithm is technically very straightforward and it is very likely that, with further development, the performance of the algorithm can be enhanced. For example, the current windowing process uses a fixed window length and step size. This forces an overlap between subsequent windows and also leads regularly to windows that span across several subsequent words. Utilizing a varying length windowing approach that is synchronized, e.g., to the temporal envelope of the speech waveform, could facilitate learning and increase model selectivity (cf., e.g., Ahissar & Ahissar,

2005). Another limitation of the current approach is that it is strictly limited to a discrete acoustic space. This makes the models susceptible to noise in the input, since the hard decisions made at the vector quantization stage lead to different representations for slightly different acoustic events. However, this is not a fundamental limitation and it is possible to extend the current algorithm to approximate a continuous acoustic space with multiple weighted VQ labels per frame, or by using a Gaussian mixture frontend instead of vector quantization. Despite the current shortcomings, the algorithm clearly demonstrates a capability for the incremental learning of internal representations from speech without supervision.

It is also noteworthy that the proposed model is purely incremental and the details of the acoustic input can be forgotten as soon as a few hundred milliseconds after the signal has arrived to the system. This also means that the model is fully unaware of the global transition probabilities of linguistic or phonetic units like phones, syllables, and words since it cannot backtrack to previously heard utterances and attempt to parse them using later learned models. Additionally, the learned word models are ready to be grounded to other information sources available to the learner, e.g., through cross-situational learning (Smith & Yu, 2008).

When compared to the PERUSE algorithm (Oates 2002) and the DTW-based approaches for unsupervised word learning (Park & Glass, 2005, 2006; Aimetti, 2009), it is evident that the previous approaches are also compatible with the idea of tracking transitional probabilities in speech. Detailed pair-wise spectral comparisons between previously perceived utterances ensure that recurring structural dependencies are detected if they exist, even though the algorithms do not explicitly manipulate the probabilities of subsequent acoustic events in a manner that is often represented in the infant learning literature (e.g., Saffran al., 1996). The first major difference between the current algorithm and the DTW-based approaches is in the manner that the memory

of the learner is organized. The DTW-based approaches and the PERUSE use full feature vector representations of all episodes, i.e., they assume acoustic episodic memory in which the heard utterances are stored. On the other hand, the algorithm proposed in this work only stores transitional probabilities between atomic acoustic events, and the actual realizations of earlier heard utterances are never revisited. Another major difference is that the current model is inherently performing temporal prediction in its processing, as the current signal is used to predict the distribution of future acoustic events. This model can also be easily extended to a word level predictor by simply replacing the low-level VQ-input with the indices of the winning model at each moment in time.

5.1 Generality of the results

Although demonstrated with speech material, the methods used in this work are not speech specific, and are therefore applicable to any type of audio or even other modalities. For example, a similar transition probability framework has been studied in automatic auditory environment segmentation and classification with promising results (Räsänen & Laine, in preparation). What this suggests is that the bootstrapping of the speech recognition process does not necessarily require special attention to phonetically and linguistically motivated features like pitch, stress, phones, or syllables. This is also supported by the finding that transition probability based segmentation of tone sequences (Saffran, Johnson, Aslin, & Newport, 1999), visual images (Kirkham, Slemmer, & Johnson, 2002) and primitive visual actions (Baldwin, Andersson, Saffran, & Meyer, 2008) has been demonstrated with human subjects. The above does not rule out that the perception of speech is adapted to the properties of the native language during the development of an infant (see, e.g., Kuhl, 2004), but simply states that the knowledge of these features is not required in advance for learning to be initiated. In later stages of

development, the plastic human brain will probably exploit any systematic properties of sensory input that exhibit a predictive value for the perceiver.

Regarding the language generality of the computational model, the experiments reported here for English speech were also performed for Finnish (CAREGIVER Y1 FIN corpus; Altosaar et al., 2010), which is a very different language from English due to its agglutinative morphology. Despite the large differences between the languages, the results were very similar and all observations drawn here for English are applicable for Finnish as well. In addition, there were some additional effects for Finnish like the learning of individual models for the most commonly occurring morphemes in inflections. On the other hand, the lack of articles in Finnish increases the average word length, possibly making the fixed window analysis more suitable for Finnish. In order to keep this work as compact as possible, the experiments performed with Finnish data were not reported in further detail.

5.2 Future work

The MFCCs used as features in this work represent the overall spectral structure that includes spectral tilt and formants. This means that cues for syllabic or phonemic identity and linguistic stress are all embedded in the single MFCC representation. In addition, timing cues are implicitly included in the model through the modeling of speech sound transitions at different temporal distances. However, in the future it would make sense to study the relative weight of different acoustic cues (e.g., energy, spectral tilt, pitch) on the segmentation performance using the proposed algorithm.

It would also be interesting to study the applicability of the proposed model to the acquisition of phone- or syllable-like segments. By using smaller analysis windows and shorter time constants in activity smoothing, and possibly synchronizing the novelty detection window

to the envelope of syllabic structure (see above), the algorithm would attempt to discover recurring structures at this smaller time scale. However, this requires the use of a much larger speech corpus, since the current vocabulary of 80 words has only a small number of overlapping syllables across words, and therefore it cannot account for the context effects of phonemes. An ideal corpus for this purpose would contain extensive amounts of speech from a limited number of talkers. Signals should be recorded in controlled environments, since additional noise outside of normal speech and talker variation make the analysis much more difficult and expensive in terms of computational complexity.

5.3 Concluding remarks

Although the current work has shown the possibility for totally unsupervised learning of proto-lexical candidates, it should not be forgotten that real linguistic development takes place in a much richer world where the learner is embedded in tight interaction with its caregivers and the surrounding environment (e.g., Meltzoff, Kuhl, Movellan, & Sejnowski, 2009). When compared to the unimodal learning situation as was used in this work, the interaction with the complex real world and other social agents actually imposes additional constraints and provides feedback that can aid in linguistic development (see, e.g., Oudeyer & Kaplan, 2006; Yu et al., 2005). Also, the only way to acquire meaning for the auditory word forms is to ground them in other perceptual systems and actions of the agent. This is something that was not studied in this work on purpose in order to see how much is possible using only a single audio source. It should also be noted that a real infant is exposed to a much larger amount of speech during infancy than what was used in this study, or any other known studies attempting to perform computational modeling of language acquisition.

Finally, the level of processing at which multimodality and feedback come into play and where grounding of words takes place, is not clear. In the study and modeling of cognitive agents, it is of great interest to understand how much can, and should be, processed in one sensory stream alone, and when is it appropriate to utilize information across several sensory modalities, motor actions, and internal states of the system in order to perform useful and efficient computations. So far, a theoretic framework for solving these types of computational learning questions is still missing.

Acknowledgements

This research was supported by Nokia NRC Tampere of the Nokia Corporation, the Finnish Graduate School in Language Studies (Langnet) funded by the Ministry of Education of Finland, and the EU FP6 FET project Acquisition of Communication and Recognition Skills (ACORNS), contract no. FP6-034362. The author is grateful to Dr. Kris Demuynck for providing the forced-alignment annotations for the CAREGIVER speech data and to Dr. Toomas Altosaar for the extensive comments and corrections to the manuscript. The author would also like to thank all three anonymous reviewers for their invaluable comments and suggestions.

References

Ahissar, E., & Ahissar, M. (2005). Processing of the temporal envelope of speech. In R. König, P. Heil, E. Budinger, & H. Scheich (Eds.): *The auditory cortex: a synthesis of human and animal research*. Lawrence Erlbaum Associates, New Jersey.

Aimetti, G. (2009). Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism. *Proceedings of EACL-2009-SRWS*, Athens, Greece, 1-9.

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Malta, 1062-1068.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition,* 106, 1382-1407.

Bortfeld, H., & Morgan, J. L. (in press). Is early word-form processing stress-full? How natural variability supports recognition. *Cognitive psychology*.

Brent, M. R. (1999a). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning,* 34, 71-105.

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

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactics are useful for segmentation. *Cognition*, 61, 93-125.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Lexical Segmentation, the role of sequential statistics in supervised and un-supervised models. *Proceedings of the 16th Annual Conference of Cognitive Science Society*, 36-141.

Ching, W. K., Fung, E. S., & Ng, M. K. (2004). High-Order Markov Chain Models for Categorical Data Sequences. *Naval Research Logistic*, 51, 557-574.

Christiansen, M. H. , Allen, J. A., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221-268.

Coen, M. H. (2006). Self-supervised acquisition of vowels in American English. *Proceedings of the 21st national conference on Artificial intelligence*, Boston, USA, 2, 451-1456.

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, 92, 81-104.

Davis S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions of Acoustics, Speech, and Signal Processing*, Vol. 28, 357-366.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands, 2208-2213.

Fenson, L., Marchman, V.A., Thal, D.J., Dale, P.S., & Bates, E. (2003). MacArthur-Bates Communicative Development Inventories (CDIs), Second Edition. Baltimore, MD: Brooks Publishing.

Gales, M., & Young, S. (2008). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1, 195-304.

Hillenbrand, J. L., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099-3111.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 1570-1582.

Huang, A. (2008). Similarity Measures for Text Document Clustering. *Proceedings of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC2008*, Christchurch, New Zealand*, 49-56.*

Huang, X. (1992). Minimizing speaker variation effects for speaker-independent speech recognition. *Proceedings of the workshop on Speech and Natural Language* of the *Human Language Technology Conference*, Harriman, New York, 191-196.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language,* 44, 548-567.

Jusczyk, P. W. (1993). Discovering sound patterns in the native language. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society*, Colorado, Boulder, 49-60.

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

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29, 1-23.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition,* 83, B35- B42.

Kuhl, P. K. (1986). Theoretical contributions of tests on animals to the special-mechanisms debate in speech. *Experimental Biology*, 45, 233-265.

Kuhl, P. K. (2004). Early Language Acquisition: Cracking the Speech Code. *Nature Reviews Neuroscience,* 5, 831-843.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Phil. Trans. Royal Society B*, 363, 979-1000.

Lake, B. M., Vallabha, G. K., & McClelland, J. L. (2009). Modeling Unsupervised Perceptual Category Learning. *IEEE Transactions on Autonomous Mental Development*, 1 35-43.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 281-297.

de Marcken, C. (1995). The unsupervised acquisition of a lexicon from continuous speech. AI Memo No. 1558, Massachusetts Institute of Technology.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12, 369-378.

Meltzoff, A. N., Kuhl, P. K., Movellan, J., & Sejnowski, T. J. (2009). Foundations for a New Science of Learning. *Science,* 325, 284-288.

Park, A., & Glass, J. R. (2005). Towards Unsupervised Pattern Discovery in Speech. *Proceedings of 2005 IEEE Workshop Automatic Speech Recognition and Understanding (ASRU'05)*, Cancún, Mexico, 53-58.

Park, A., & Glass, J. R. (2006). Unsupervised word acquisition from speech using pattern discovery. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 409-412.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Eight-month-old infants track backward transitional probabilities. *Cognition*, 113, 244-247.

Pisoni, D. B. (1997). Some Thoughts on "Normalization" in Speech Perception. In K. Johnson and J. W. Mullennix (Eds.): *Talker Variability in Speech Processing*, San Diego: Academic Press, 9-32.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology,* 25, 143-170.

Oates, T. (2001). *Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning*. Doctoral Thesis, University of Massachusetts Amherst, USA.

Oates, T. (2002). PERUSE: An unsupervised algorithm for finding recurrent patterns in time-series. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, Maebashi City, Japan, 330-337.

Oudeyer, P.-Y., & Kaplan, F. (2006). Discovering Communication. *Connection Science*, 18, 189-206.

Raftery, A. E. (1985). A new model for discrete-valued time series: Autocorrelations and extensions. *Rassegna di Metodi Statistici ed Applicazioni*, 3-4, 149-162.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica,* 14, 465-471.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12, 339-349.

Räsänen, O. J., & Laine, U. K. (in review). A method for noise robust context-aware pattern discovery from symbolic time series.

Räsänen, O. J., Laine, U. K., & Altosaar, T. (in press). Blind segmentation of speech using non-linear filtering methods. In I. Ipsic (Ed.): *Speech Technologies.* Accepted book chapter.

Räsänen, O. J., Laine, U. K., & Altosaar, T. (2009a). A noise robust method for pattern discovery in quantized time series: the concept matrix approach. *Proceedings of 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, Brighton, England, 3035-3038.

Räsänen, O. J., Laine, U. K., & Altosaar, T. (2009b). Self-learning Vector Quantization for Pattern Discovery from Speech. *Proceedings of 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, Brighton, England, 852-855.

Räsänen, O. J., Laine, U. K., & Altosaar, T. (2008). Computational language acquisition by statistical bottom-up processing. *Proceedings of 9th Annual Conference of the International Speech Communication Association (Interspeech'09)*, Brisbane, Australia, 1980-1983.

Saffran, J. R. (2001). Words in the sea of sounds: the output of infant statistical learning. *Cognition,* 81, 149-169.

Saffran, J. R., Newport, E. L., & Aslin R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35, 606-621.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, 274, 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition,* 70, 27-52.

Scharenborg, O., Ernestus, M., & Wan V. (2007). Segmentation of speech: Child's play? *Proceedings of 8th Annual Conference of the International Speech Communication Association (Interspeech'07)*, Antwerp, Belgium, 1953-1956.

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modeling of auditory word perception in continuous speech. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* Bloomington, Indiana, 408-413.

Singh, L., White, K. S., & Morgan J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4, 157-178.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558-1568.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86-132.

Thiessen, E., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39, 706-716.

Thiessen, E., & Saffran, J. R. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics*, 65, 779-791.

Toledano, D. T., Hernández Gómez, L. A., & Villarrubia Grande, L. (2003). Automatic Phonetic Segmentation. *IEEE Transactions on Speech and Audio Processing*, 11, 617-625.

Toscano, J. C., & McMurray, B. (2010). Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics. *Cognitive Science*, 34, 434-464.

Vallabha, G. K., McLelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of National Academy of Sciences*, 104, 13273-13278.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27, 351-372.

Warren, R. M. (2000). Phonemic organization does not occur: Hence no feedback. Commentary to Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 350-351.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, 1, 197-234.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence from perceptual reorganization during the first year of life. *Infant Behavior and Development,* 7, 49-63.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59, 114-132.

Yang, C. D. (2004). Universal Grammar, statistics, or both? *TRENDS in Cognitive Sciences*, 8, 451-456.

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The Role of Embodied Intention in Early Lexical Acquisition. *Cognitive Science,* 29, 961-1005.

Appendix A

Table A1: Vocabulary of the CAREGIVER Y2 UK corpus

'#h'	'cookie'	'here'	'sells'
'<sil>'	'cow'	'horse'	'she'
'a'	'crying'	'i'	'small'
'airplane'	'daddy'	'is'	'smiling'
'an'	'dirty'	'like'	'square'
'and'	'do'	'likes'	'stares'
'animal'	'dog'	'lion'	'take'
'apple'	'doll'	'looks'	'takes'
'at'	'duck'	'man'	'telephone'
'baby'	'eagle'	'mean'	'that'
'ball'	'edible'	'mummy'	'the'
'banana'	'fish'	'no'	'there'
'big'	'frog'	'porsche'	'this'
'bird'	'give'	'red'	'toy'
'blue'	'gives'	'robin'	'tree'
'bottle'	'had'	'round'	'truck'
'car'	'happy'	'sad'	'where'
'cares'	'has'	'said'	'woman'
'cat'	'have'	'see'	'yellow'
'clean'	'he'	'sees'	'you'

Appendix B

Computation of MFCCs: The incoming speech signal *x* is windowed into a series of frames using a Hamming window function *w*

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad \text{(B1)}$$

where *N* is the total length of the window. In other words, *N* subsequent samples of the waveform are chosen and each sample *x*[n], $n = \{1, \ldots, N\}$, is multiplied by the Hamming window value *w*[n].

$$y[n] = x[n]w[n] \quad \text{(B2)}$$

In this work, a window length of 32 ms (512 samples at the sampling rate of 16 kHz) was used. Next, the fast Fourier Transform (FFT) is applied to the windowed signal ***y*** to obtain the power spectrum of the signal:

$$X[k] = abs\left[\sum_{n=0}^{N-1} y[n]e^{-i2\pi k\frac{n}{N}}\right], \quad k = \{1,...,N\} \quad \text{(B3)}$$

Since the human ear has higher frequency resolution at lower frequencies, a Mel-scale filterbank is applied to the FFT-spectrum to simulate this effect. The Mel-filterbank in the experiments of this study consisted of 26 triangular bandpass filters whose center frequencies and bandwidths increase logarithmically as a function of frequency (Figure C1). The filterbank was built by first computing the maximum Mel-frequency given the signal sample rate by having $Mel_{\max} = 2595*\log_{10}(1+f_s/(2*700))$, where f_s is the sampling rate, and then by dividing the resulting Mel-range [0, …, $Mel_{\max}$] into 26 uniformly spaced filter center frequencies in the Mel domain. These center frequencies are converted back to the frequency domain by having $f = 700(10^{m/2595} - 1)$ where *m* is the center Mel-value of each band. Each triangular band starts at

the center frequency of the neighboring lower band and ends at the center frequency of higher frequency band.

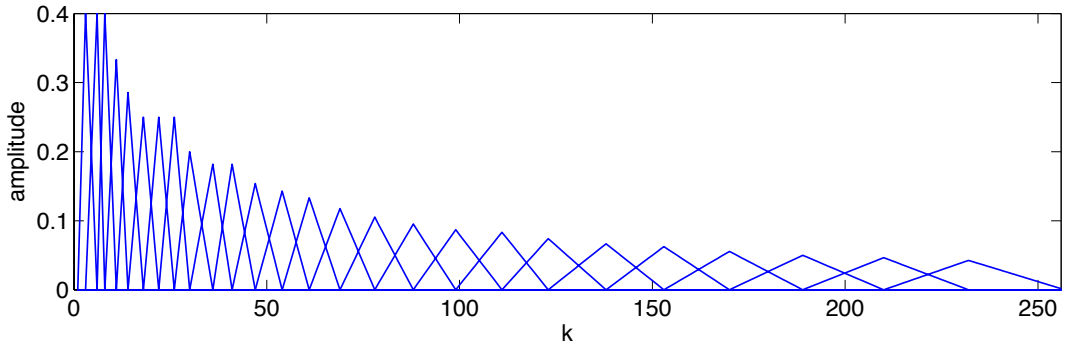


Fig. C1. Mel-filterbank applied to the FFT-spectrum in the computation of Mel-spectrum.

The FFT power spectrum is then filtered (multiplied in the frequency domain) using the Mel-scale filterbank in order to obtain the Mel-spectrum of the windowed speech signal. The logarithm of the power in each band is taken to produce the Mel-log spectrum ***f'***. Finally, the discrete cosine transform is applied to the mel-log spectrum to obtain the Mel-frequency cepstral coefficients (MFCCs):

$$c_k = \sum_{n=1}^{N} f'[n]\cos\left[\frac{\pi}{N-1}(n+0.5)k\right] \quad k = \{0,...,N-1\} \qquad \text{(B4)}$$

Appendix C

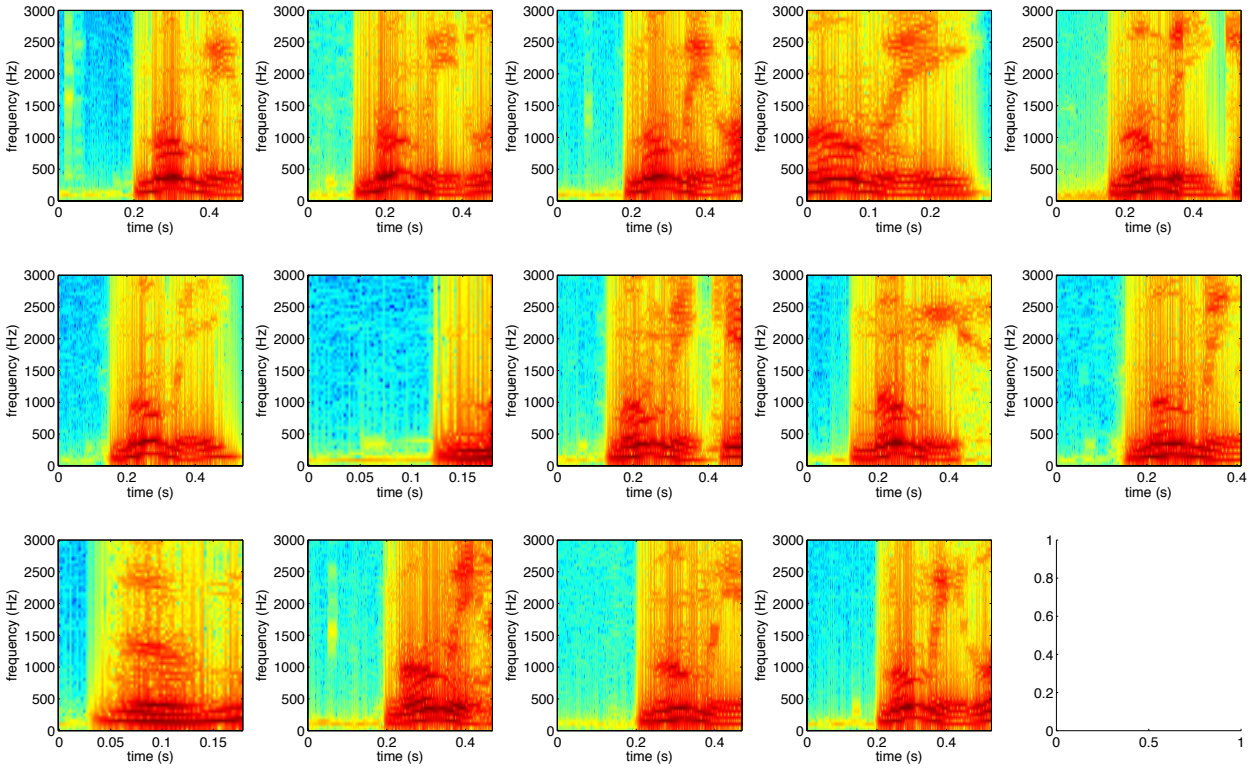
Table C1: Contents of the word models in the high-threshold condition sorted according to model entropy.

#	N	α_1	p	α_2	p	α_3	p	#	N	α_1	p	α_2	p	α_3	p
155	2	frog	1.00					162	5	yellow	0.66	truck	0.19	lion	0.04
172	2	baby	1.00					200	10	baby	0.61	sees	0.23	takes	0.06
15	91	#h	1.00					61	9	mummy	0.63	has	0.15	#h	0.12
92	91	#h	0.99	lion	0.01			122	10	square	0.73	cat	0.05	frog	0.05
150	3	red	0.98	the	0.02			180	6	yellow	0.54	horse	0.33	round	0.04
31	2	airplane	0.98	#h	0.02			152	6	bottle	0.53	robin	0.29	#h	0.09
2	104	#h	0.96	porsche	0.02	bottle	0.01	28	10	cookie	0.63	#h	0.20	looks	0.06
128	2	bird	0.94	and	0.06			49	11	sad	0.72	the	0.07	woman	0.04
21	14	telephone	0.95	and	0.02	happy	0.01	77	4	clean	0.54	likes	0.24	a	0.16
126	32	#h	0.93	dog	0.02	duck	0.02	117	6	here	0.55	is	0.25	#h	0.11
116	5	airplane	0.91	and	0.05	a	0.03	188	5	has	0.57	daddy	0.26	baby	0.06
5	19	#h	0.92	eagle	0.02	doll	0.02	69	4	apple	0.50	dog	0.23	toy	0.21
8	55	#h	0.92	doll	0.02	there	0.02	96	6	sad	0.69	the	0.09	animal	0.04
89	4	sad	0.89	the	0.05	a	0.03	53	8	banana	0.49	blue	0.34	car	0.09
62	15	#h	0.89	frog	0.05	dog	0.04	95	5	gives	0.61	man	0.16	a	0.08
169	5	said	0.87	small	0.06	square	0.04	80	7	baby	0.43	likes	0.34	he	0.11
42	6	dirty	0.87	the	0.07	is	0.02	6	14	mummy	0.57	#h	0.26	the	0.06
107	6	bird	0.87	#h	0.04	a	0.04	204	3	red	0.42	the	0.26	dirty	0.18
195	4	edible	0.82	the	0.15	sees	0.03	108	9	mean	0.64	i	0.16	no	0.18
63	6	duck	0.87	red	0.04	and	0.04	223	3	daddy	0.33	sees	0.33	baby	0.24
10	7	happy	0.84	the	0.11	gives	0.02	67	9	bird	0.51	#h	0.23	car	0.16
84	6	cookie	0.85	square	0.07	and	0.03	90	3	happy	0.54	woman	0.16	a	0.15
18	39	#h	0.82	horse	0.15	looks	0.03	11	6	square	0.58	small	0.19	a	0.07
121	9	airplane	0.78	#h	0.20	happy	0.01	129	6	gives	0.52	square	0.29	a	0.06
151	12	banana	0.76	#h	0.22	and	0.02	60	23	no	0.57	#h	0.14	i	0.14
192	9	bottle	0.85	#h	0.06	and	0.04	43	12	looks	0.62	at	0.17	daddy	0.05
137	8	round	0.83	gives	0.09	a	0.07	100	19	frog	0.52	#h	0.19	sees	0.17
48	56	#h	0.80	fish	0.17	porsche	0.02	65	2	sees	0.50	yellow	0.24	is	0.11
45	8	#h	0.75	lion	0.23	and	0.02	213	5	#h	0.41	truck	0.26	big	0.24
132	6	#h	0.71	robin	0.26	the	0.02	160	2	telephone	0.50	and	0.19	apple	0.17
211	3	telephone	0.79	big	0.12	red	0.05	23	18	clean	0.56	car	0.29	horse	0.02
149	2	telephone	0.72	dirty	0.21	big	0.07	134	7	round	0.62	the	0.12	red	0.08
167	3	apple	0.76	and	0.13	happy	0.07	3	2	happy	0.53	the	0.17	there	0.12
181	3	daddy	0.72	#h	0.19	gives	0.07	58	3	looks	0.43	she	0.32	#h	0.10
7	12	animal	0.71	#h	0.24	an	0.03	170	27	dog	0.43	doll	0.40	dirty	0.05
55	4	bird	0.71	dirty	0.22	round	0.04	91	10	here	0.43	#h	0.23	is	0.20
85	11	toy	0.77	#h	0.15	clean	0.02	32	14	cow	0.45	cat	0.31	#h	0.12
193	4	happy	0.67	woman	0.25	mummy	0.04	35	16	smiling	0.54	sad	0.17	#h	0.14
26	14	sees	0.76	baby	0.09	she	0.08	79	10	square	0.65	#h	0.05	a	0.05
119	3	woman	0.66	sees	0.19	takes	0.15	157	5	likes	0.47	she	0.26	#h	0.09
64	4	daddy	0.77	takes	0.06	#h	0.06	66	4	she	0.32	has	0.31	#h	0.22
39	9	dirty	0.78	is	0.07	a	0.06	83	38	#h	0.48	eagle	0.19	apple	0.15
184	3	edible	0.63	toy	0.21	cookie	0.16	17	13	where	0.54	is	0.13	the	0.10
218	2	has	0.62	daddy	0.24	the	0.14	165	3	eagle	0.43	sees	0.20	the	0.12
99	6	happy	0.77	dirty	0.06	sees	0.05	54	7	here	0.31	there	0.26	is	0.21
13	6	likes	0.69	man	0.20	woman	0.06	47	12	crying	0.58	the	0.13	woman	0.07
110	9	man	0.74	takes	0.11	gives	0.06	38	6	frog	0.46	airplane	0.16	red	0.15
74	6	likes	0.75	daddy	0.07	the	0.06	16	14	tree	0.44	small	0.36	#h	0.05
59	15	edible	0.76	the	0.09	animal	0.07	36	17	eagle	0.61	#h	0.13	an	0.05
147	5	gives	0.62	woman	0.26	the	0.07	123	8	sees	0.55	daddy	0.15	blue	0.07
130	12	car	0.72	#h	0.18	a	0.03	4	15	porsche	0.36	ball	0.29	horse	0.23
118	2	dog	0.50	doll	0.34	sad	0.16	75	4	robin	0.47	square	0.15	is	0.12
12	11	small	0.79	a	0.04	the	0.03	14	7	likes	0.55	the	0.11	yellow	0.10
34	19	#h	0.59	toy	0.29	doll	0.08	228	4	takes	0.39	she	0.26	big	0.17
88	7	duck	0.72	dog	0.14	a	0.04	109	20	truck	0.56	tree	0.15	dirty	0.06
140	2	yellow	0.71	apple	0.08	the	0.08	222	4	daddy	0.25	has	0.25	mummy	0.22
205	9	edible	0.75	woman	0.07	bottle	0.06	86	6	gives	0.29	baby	0.28	she	0.20
158	10	#h	0.51	lion	0.41	yellow	0.03	120	10	clean	0.40	frog	0.26	truck	0.15
70	4	no	0.50	animal	0.25	#h	0.24	46	6	happy	0.43	gives	0.21	has	0.11
94	9	woman	0.71	the	0.09	a	0.08	40	16	like	0.35	you	0.21	do	0.18
27	3	sees	0.55	dirty	0.33	an	0.07	104	4	looks	0.40	big	0.18	at	0.15
72	4	blue	0.70	big	0.09	toy	0.09	102	8	cow	0.35	cat	0.22	#h	0.19
25	18	happy	0.73	the	0.11	sees	0.06	93	7	crying	0.33	round	0.29	happy	0.13
9	16	takes	0.76	a	0.05	big	0.04	22	25	where	0.43	#h	0.15	is	0.13
240	2	#h	0.47	lion	0.38	square	0.13	20	22	ball	0.45	porsche	0.16	small	0.15
135	28	#h	0.59	tree	0.30	baby	0.04	76	7	have	0.33	round	0.19	you	0.15
50	4	looks	0.47	man	0.42	at	0.07	133	13	like	0.36	square	0.18	do	0.10
136	10	blue	0.74	a	0.07	the	0.05	68	3	have	0.29	the	0.18	you	0.16
52	10	lion	0.68	#h	0.17	and	0.06	101	10	red	0.31	round	0.19	is	0.13
103	10	cow	0.53	cat	0.37	a	0.03	106	17	#h	0.22	banana	0.18	you	0.18
201	7	frog	0.65	blue	0.19	clean	0.06	19	15	man	0.34	smiling	0.22	round	0.10
164	12	#h	0.70	smiling	0.08	lion	0.07	125	12	have	0.35	you	0.16	#h	0.14
141	4	#h	0.44	there	0.39	is	0.14	30	7	animal	0.27	horse	0.14	apple	0.14
124	2	like	0.58	you	0.25	blue	0.10	159	3	see	0.19	happy	0.15	the	0.15
								57	29	see	0.26	you	0.21	#h	0.13

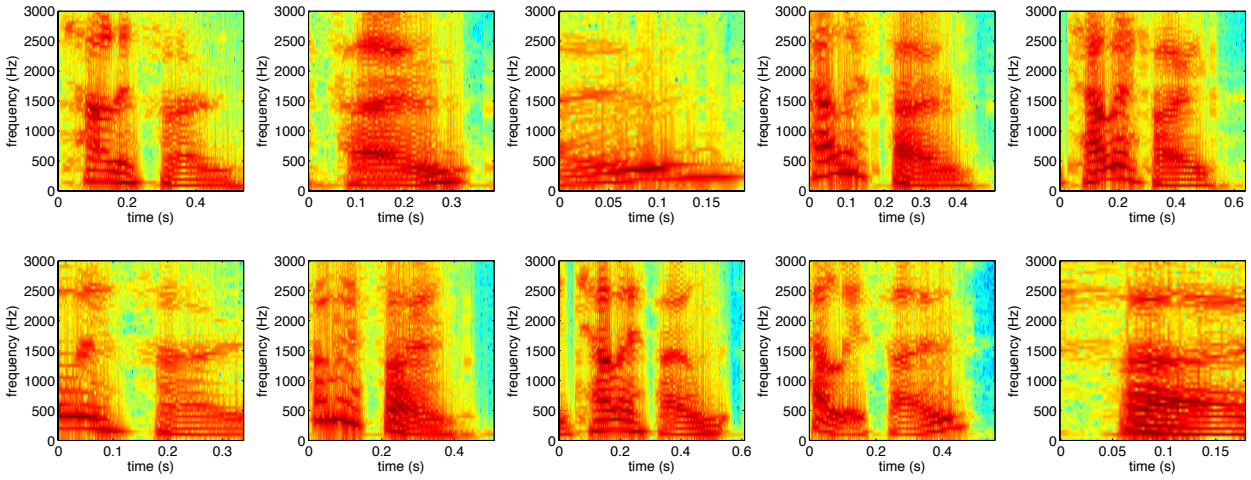
Appendix D: Examples of automatically segmented words

Realizations from different models are shown below. Realizations are extracted automatically in the order of appearance in the test set. The models are named by their most dominant word, although they may contain a number of deviating patterns.

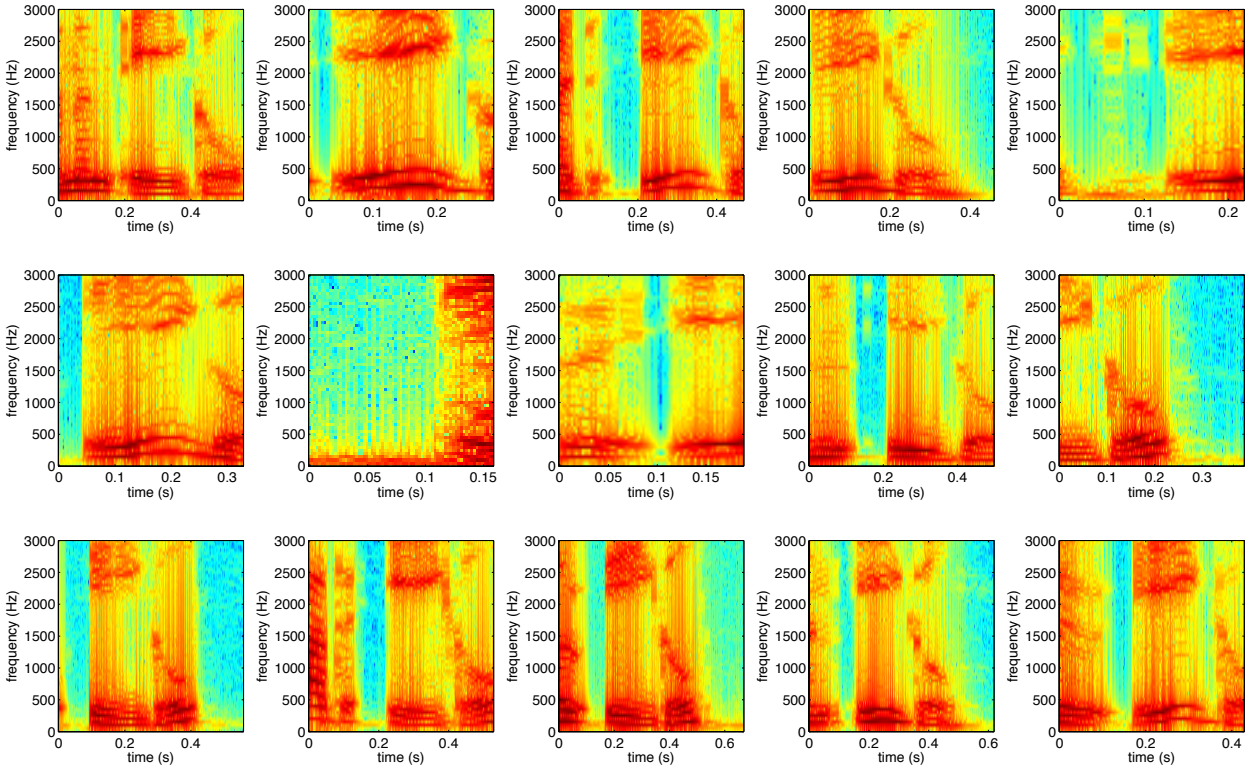
"Mommy"



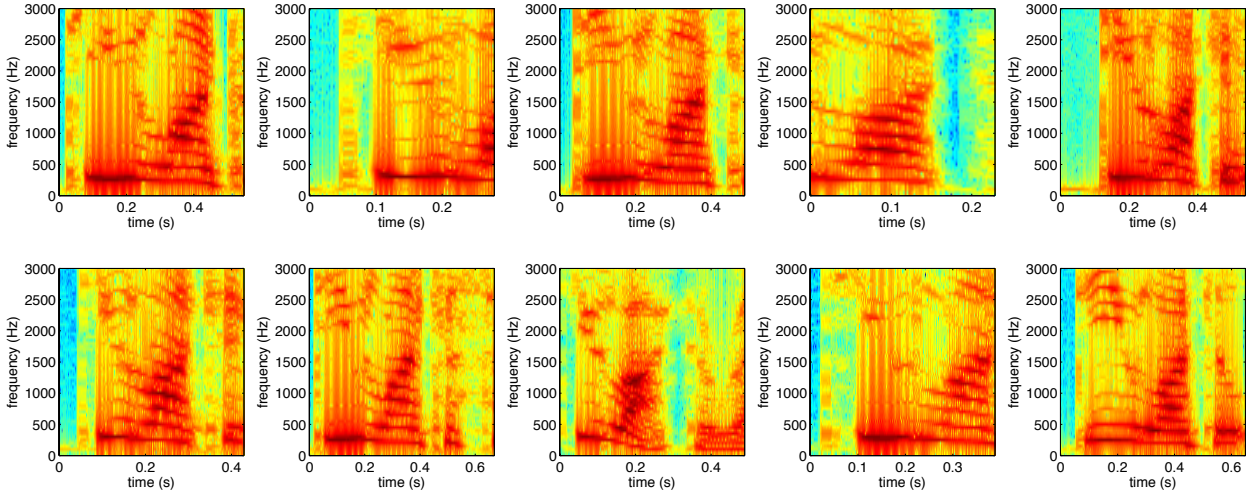
"Telephone" (sometimes extracted as *"phone"*).



"Eagle"



"Doyoulikethe"



"Thecrying" (man/woman)