Cognition 120 (2011) 149–176

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

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

ARTICLE INFO

Article history:
Received 8 June 2010
Revised 31 March 2011
Accepted 2 April 2011
Available online 27 April 2011

Keywords:
Unsupervised learning
Language acquisition
Word segmentation
Distributional learning

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.

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

* Tel.: +358 9 470 22499; fax: +358 9 460 224.
E-mail address: okko.rasanen@aalto.fi

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.04.001

a language. In his work, Jusczyk (1993) has shown that 6-months-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-months-old infants prefer to use transitional probabilities of speech sounds in segmentation whereas 9-months-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 (Bortfeld & Morgan, 2010; Houston & Jusczyk, 2000; 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-months-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