Similarities in fundamental frequency in infant speech segmentation models

Ellen Marklund, Francisco Lacerda, Iris-Corinna Schwarz, Ulla Sundberg

Department of Linguistics, Stockholm University, Stockholm, Sweden
ellen@ling.su.se, frasse@ling.su.se, iris@ling.su.se, ulla@ling.su.se

Abstract

The present study investigates fundamental frequency as a potential basis for segmentation in models of infant speech segmentation. Pairs of segments that were similar either in terms of fundamental frequency envelop or in terms of transcribed content were found in three different speech styles: speech directed to three-month-olds, speech directed to twelve-month-olds and speech directed to adults. Spectral distance between the segments was calculated for each pair and used as a rough measure of spectral similarity. In both the infant-directed speech style conditions, the spectral distance was smaller when fundamental frequency was used as basis for segmentation, compared to when transcriptions were used. In the adult-directed speech style condition, no difference was found between different bases of segmentation.

Index Terms: speech segmentation, fundamental frequency, language acquisition, infant-directed speech

1. Background

The interest in modeling human language acquisition has recently surged. To model the developing linguistic capacities of infants in an empirically plausible way is both an interesting challenge in itself, as well as way to test hypotheses and theories about human cognition. One of the tasks that infants are faced with when acquiring their first language is to figure out where word boundaries are. Adopting an ecological view of language acquisition, it is required that both the infants’ capacities and their linguistic environment change as they develop rapidly during the first years of life. A model with the purpose of modeling infant speech segmentation need either specify the intended age of the modeled infant, or implement developmental changes over time. Below, overviews of infants’ speech segmentation abilities and linguistic input are given, followed by a presentation of the present study.

1.1. Speech segmentation

Already at birth, infants possess the perceptual capacities necessary to detect word segmentation cues [1]. Their abilities to segment single words from continuous speech streams have been investigated regarding several different types of cues. One example is prosody; infants are able to segment words with a more common stress pattern from continuous speech at an earlier age than they are able to segment words with a more infrequent stress pattern [e.g. 2], suggesting that they make use of the prosodic information when segmenting speech. Another cue infants utilize is the silence preceding and following utterances (which by extension also precedes or follows the first and last word of the utterances). Infants have been shown to be better at segmenting both utterance-initial and utterance-final words from a continuous speech than they were on segmenting words in utterance-medial position [3]. Transitional probabilities are another cue to which infants have been reported to be sensitive, both in artificial speech [e.g. 4] and natural language [e.g. 5]. Infants’ reliance on different word boundary cues varies as they mature and as they acquire more experience with the native language. For example, 9-month-olds are sensitive to language-specific phonotactic word boundary cues [6], whereas sensitivity to allophonic cues has not been reported until 10.5 months of age [7].

When modeling infant word segmentation, different strategies can be employed depending on possible theoretical assumptions about the lexical representations of the infants, as well as the age and developmental stage of the infant. Since infants are able to recognize familiar words already at a very young age [8], and several studies have indicated that infants have a fairly detailed lexical representation of words they know [e.g. 9], the segmentation models mentioned in the present paper will be the word-recognition type, as described by Brent [10]. The word-recognition strategy entails storing input as lexical candidates (hypothesized words), and recognizing those stored units in future input. Adopting an ecological view of language acquisition [e.g. 11], an analytic approach is preferable to a synthetic approach, since the former eliminates the need for a priori linguistic knowledge. A number of models with the explicitly stated intent of modeling infant speech segmentation match these criteria, e.g. PUDDLE [12], the LA-model [13] and the emergent lexicon model [14].

1.2. Speech input

Adults modify their speech style when talking to infants. Infant-directed speech (IDS) is characterized by highly modified prosodic patterns [e.g. 15], linguistic simplifications [e.g. 16, 17], temporal modifications [e.g. 18], as well as modifications on a segmental level, such as expanded vowel space [e.g. 19] and modified voice onset time [e.g. 20]. The IDS style has been reported cross-linguistically [e.g. 21], and regardless of speaker gender [22]. The characteristics of IDS are more or less pronounced depending both on the characteristics of the acquired language and on the developmental stage of the infant the speech is directed to [e.g. 23].

About half of the speech one infant was exposed to during a day was infant-directed or child-directed speech [24]. Additionally, infants prefer listening to IDS over adult-directed speech [e.g. 25], regardless of speaker gender [26] and language [27]. Both those things suggest that infant-directed speech is a salient form of input to infants, although the specific characteristics and modifications vary with language of the speaker and infant age. Therefore, in order to model infant word segmentation, the speech input should preferably be infant directed speech, since this is a salient aspect of what the infant is
exposed to. Further, the content and form of the speech input would need to differ depending on the intended “age” of the model. Furthermore, the segmentation should be done on speech represented by auditory information, using segmenting criteria that match the auditory and cognitive capacities of infants at the target age. The three models mentioned above all use IDS (or child-directed speech, CDS) or a simulation thereof. PUDDLE uses phonematic transcriptions of English CDS from natural adult-infant interactions [12], and the emergent lexicon model uses natural transcribed Swedish IDS [14]. In contrast, the LA-model uses a representation of the actual speech signal, and in the report from 2009 the sentences were content-wise simulations of IDS in that they were simple questions of the kind “Do you see the [target word]?” [13].

1.3. The present study

The present study investigates the possibility of using repetitions in fundamental frequency contours as basis for segmentation in models intended to simulate speech segmentation at a young age. A word-recognition segmentation procedure with an analytic approach will be used, since it assumes no prior linguistic knowledge.

Fundamental frequency is used because it is a salient aspect of IDS, with higher variation and an overall higher range compared to adult-directed speech (ADS), with greatest variations when the infants are four months old [18]. IDS also contains a large number of prosodic repetitions, sometimes co-occurring with semantic repetitions [15]. Four-month-olds were tested on their preference for manipulated IDS, in which different characteristics had been isolated (e.g. fundamental frequency and amplitude), and showed a preference for the simulated IDS only in the fundamental frequency condition [28], suggesting that fundamental frequency is the aspect of IDS infants at this age find to be most attractive. Consequently, the speech used in the present study was IDS directed to infants at about this age. Samples of IDS towards children at about one year of age, as well as ADS, were also included for comparison. However, older infants (and certainly adults) are already linguistically proficient, and most likely use other cues for speech segmentation.

Segment pairs for spectral comparison were found in the speech input by highlighting similar segments of fundamental frequency. In order to have reference points on the current scale of spectral distance, similar segments based on transcriptions were also found and analyzed. Additionally, pairs at quasi-randomly chosen places in the speech material (matching the transcription-based pairs for utterance of origin and duration) will be included in the analysis as reference points. The spectral comparison is considered a rough measure of acoustic similarity; the shorter spectral distance (measured as the accumulated difference at 40 data points along the two segments), the more similar the segments are in terms of spectral information.

To summarize, three types of input were tested using three different bases for segmentation. Among these, the one combination that can be considered empirically plausible in combination with the segmentation strategy used is IDS directed to young infants and fundamental frequency as basis for segmentation. This combination is expected to yield the smallest spectral distance, i.e. finding segments that are more similar in terms of spectral information compared to the other combinations.

2. Method

Utterances from IDS (at two different ages) and ADS were extracted and compared based on similarity, using fundamental frequency and orthographic transcriptions as a basis for comparison. Similar segments were flagged and paired. Additionally, random segment pairs were selected. The spectral distance between each pair was calculated, measured as the accumulated difference between the first three formants at three evenly spaced points in the segments. The mean spectral distance was compared between groups.

2.1. Speech material

Three types of speech styles were used in the present study; speech directed to three-month-olds (IDS03), speech directed to twelve-month-olds (IDS12), and ADS. Speakers were three female native speakers of Swedish with infants at the age of three or twelve months at the time of the recording. The IDS was recorded in a sound-attenuated room, in which the mother and infant were placed face-to-face, and the mother was instructed to play with her infant as she would at home. The ADS was elicited by an experimenter entering the room and talking informally with the mother for a few minutes.

Two of the speakers contributed to the IDS03 and the ADS condition while the third speakers’ speech was used in the IDS12 condition. The recordings were divided into 26 shorter sections, ten in the IDS03 condition and eight each in the IDS12 and ADS conditions. The total duration of the material used in the IDS03 condition was 11 min 58 s, in the IDS12 condition it was 16 min 50 s, and in the ADS condition it was 10 min 29 s.

Within the selected material, on- and offset for each utterance was marked, and the utterance was orthographically transcribed. Utterances that were clearly infant-directed were not marked in the ADS, and vice versa. Fundamental frequency data and formant data was extracted using Wavesurfer 1.8.5 [29]. The fundamental frequency data was converted from Hertz to cents with 110 Hz as reference (1), and the formant data was converted from Hertz to Bark [30], see (2).

\[ c = \frac{1200 \log(f)}{\log(110)} \]  \hspace{1cm} (1)

\[ f = \text{Frequency in Hertz, } c = \text{cent} \]

\[ b = \frac{26.81f}{2960 + f} - 0.53 \]  \hspace{1cm} (2)

\[ f = \text{Frequency in Hertz, } b = \text{Bark} \]

2.2. Segment comparison

The marked up utterances were segmented using a modified version of the dpn-gram technique [13]. In this technique, two strings of sequential data are compared; the difference (or distance) is calculated between each data point in the first string of data and all data points in the second string, resulting in a matrix of distances between all combinations of data points. Each cell of the distance matrix is given an accumulated score based on the distance values of the surrounding cells (see [31] for more details on how the scores were set in the present study). Additionally, a backtracking pointer is set for each cell to indicate which of the backwards-surrounding cells contains the lowest value. Finally, the cell with the highest score is found
(signaling offset of the longest similar segment found) and the backtracking pointers are followed until the cell’s score reaches a cut-off value (signaling the onset of the same longest similar segment found). Only the longest similar segment was highlighted.

All utterances within each recording section were run through the script highlighting the longest similar segments. In the fundamental frequency condition, each sample in the fundamental frequency data file constituted a data point. For whispered or unvoiced passages (e.g. those lacking fundamental frequency information), the distance was set to Null, and the cells were scored to accommodate this kind of noisy data. The 25 segment pairs with the longest average duration were selected for analysis. In the transcription condition, each character in the transcribed utterance constituted a data point and the values in the distance matrix were not differences but a binary indicator (same or different character). The 25 pairs with the longest matching character strings were selected for analysis. Additionally, 25 randomly selected segment pairs were created, matching the transcription segment pairs in terms of duration and utterance of origin.

In order to compare the segment pairs a rough acoustic measure of spectral distance was used; the accumulated difference in Bark between the two segments at 40 data points (the first four formants at ten equidistant points along the time-axis). The mean spectral distance was then calculated and compared between the different bases for segmentation for all speech type conditions.

3. Results

One-way ANOVAs were performed on each speech type category, and a simple contrast test was performed using the transcription condition as reference. For IDS03, a significant difference was found between the different bases for segmentation (F(2,47)=9.679, p < 0.001), and the contrast test revealed a significant difference between the transcription-based segments and the segments based on fundamental frequency (p < 0.001), but not between transcription based segment pairs and randomly selected segment pairs (figure 1, left). Also for IDS12 a significant difference between the different bases for segmentation was found (F(2,74)=3.838, p < 0.026). The contrast test showed a difference between the transcription condition and the fundamental frequency condition (p < 0.007), but no difference between the transcription condition and the random condition (figure 1, middle). For ADS, no difference was found between the different segmentation conditions (F(2,74)=0.257, p < 0.774), see figure 1, right. No comparisons were made between speech style conditions, due to the unbalanced distribution of speakers.

4. Discussion

There was no difference in mean spectral distance between the transcription based segment pairs and the randomly assigned segment pairs in either of the speech type conditions. In both of the IDS conditions, the mean spectral distance of fundamental frequency based segment pairs differed significantly from the transcription based segment pairs, while in the ADS condition it did not. This indicates that for models of infant speech segmentation, fundamental frequency may be a viable source of information for similarity-based speech segmentation.

There are of course a number of additional aspects of modeling infant speech segmentation. One example is the necessity of some sort of memory simulation. In the present segmentation process, a very crude memory simulation was used by dividing the recorded conversations into shorter sections, and comparing every utterance within a section only to all other utterance within the same section. This could be described as simulating perfect recall within those shorter sections, and total memory loss between them. Using incremental processing instead of batch processing would constitute a more nuanced memory simulation.
It has been proposed that the characteristics of IDS facilitate different aspects of the language learning process, including word segmentation by marking word boundaries. One example is that infants perform better in word segmentation tasks when the target word is presented in an utterance-final position [32, 33], and novel and focused words are often presented in utterance-final position [34]. Additionally, IDS has a relatively large number of short pauses compared to ADS [14] which has the effect of increasing the number of words in the salient utterance-final position by increasing the number of explicitly marked utterance-boundaries. The prosodic properties of IDS have also been shown to enhance speech segmentation; infants were exposed nonsense sentences with the intonational properties of either IDS or ADS, and only infants in the IDS condition were able to segment the target words from the speech stream, despite the fact that the only cue to word boundaries in both conditions was the probability with which the syllables within the nonsense sentence co-occurred [35].

In summary, the results of the present study also point towards fundamental frequency as a relevant source of information for speech segmentation in infants and the modeling thereof, considering that the smallest spectral distance (a rough measure of the similarity of two segments) was found in the IDS conditions, using fundamental frequency as basis for segmentation.

5. References