Obtaining linguistic structure from continuous speech

Svärd, N.

journal: Proceedings of Fonetik, TMH-QPSR
volume: 44
number: 1
year: 2002
pages: 009-012

http://www.speech.kth.se/qpsr
Obtaining linguistic structure in continuous speech

Nina Svärd, Pia Nehme and Francisco Lacerda,
Dept. of Linguistics, Stockholm University

Abstract
A simple model that structures continuous speech signals by suggesting word candidates is presented in this paper. To create possible representations of target word candidates a program was written in C++. Text materials in Portuguese and “Svensiska”, a nonsense language containing similar phonotactic and prosodic features as Swedish were used in this study. Since the model aims to resemble the language learning process of the infant, the material was based on infant directed speech.

The results showed that certain words in the Portuguese material were successfully extracted as being word candidates by the model. However, the results indicated that there were no such patterns to be found in the Svensiska material, which may be attributed to the structure of the Svensiska material. The importance of the relative frequency of sound sequences, their position within the phrase and the variety in the speech signal to the infants’ ability to discover patterns and words in the speech signal are discussed.

Introduction
The purpose of this study is to create and evaluate a simple model which can structure continuous speech signals. The speech sound sequences consisting of infant directed speech represents different input as heard by the infant. Infant directed speech is characterised by a high-pitched and modulated melody with many pauses and repeatings (Sundberg, 1998). These sequences, which in fact are representations of continuous/multimodal speech were in this study chosen to be transcribed orthographically.

The model aims to resemble the early language acquisition of the infant, where the plasticity of the infants’ memory and the limitations of the memory are crucial for utterances to be stored correctly (Elman 1999). On the basis of the relative frequencies of sound strings of different lengths, the model shall suggest probable word candidates (or utterance candidates), which are prominent in the material. In difference to the models which have been reviewed in literature and which have been based on an already acquired information stored in memory, this model has no stored information. Our starting point is on a very basic level where we make assumptions regarding the surroundings of the infant and not so much regarding the infant itself. The child is seen as a system of plasticity with a memory which is being structured in accordance with its surroundings.

In order for the infant to be able to perceive and discriminate sound sequences in a continuous speech flow it is assumed they also can discriminate isolated sound sequences. Perception studies have shown that infants generally can discriminate between all speech sounds which they are exposed to (Aslin, Jusczyk & Pisoni 1998). However, at the end of the first year a change takes place so that the ability to discriminate starts to resembles that of the adult in the same language community (Werker & Tees 1984; Kuhl, Williams, Lacerda, Stevens & Lindblom 1992). This leads to the infants ceasing to discriminate between sound contrast they earlier have been able to discriminate between (Werker & Tees, 1984; Best, McRoberts, LaFleur & Silver-Isenstadt, 1995; Polka & Werker 1994). This can be seen as a cognitive development, where the children instead rely on the acoustic dimensions important to the mother tongue and filters irrelevant information.

---
1 With utterances we here mean the units of speech which are delimited with easily recognisable acoustic borders and which often consist of multiple words.
The infant is exposed to the speech signal which carries variations and this can be seen as an asset and as being of a decisive importance for its contingency to extract underlying linguistic principles (Lacerda & Sundberg, 2001). Through connecting different sensory impressions with the speech signal the infant acquire linguistic structure. The speech signal tends to systematic be related to other sensory input, e.g. objects in the near surroundings of the infant since the spoken language is used exactly for referring to objects and occurrences in the immediate surrounding of the infant.

Method

Input
The material consisted of infant directed speech in Portuguese and in "Svensiska", a nonsense language sharing similar phonotactic and prosodic features with Swedish. To resemble a continuous speech signal the material was transcribed as one single string and only natural pauses were marked. The choice of material is justified as infant directed speech contains a profitable structure for the infant in the learning process. In its transcribed version the Portuguese material consisted of 55 utterances making out a total of 710 symbols. The Svensiska material consisted of 66 utterances making out a total of 940 symbols.

Program
In order to obtain possible representations of target word candidates, a program was written in C++. The program takes as input a sequence of speech sound symbols and pause symbols $F$ and a length of sequence $L$. The output consists of the frequencies for each unique sequence $S$ by length $L$ which is in $F$ and that moreover contains a symbol marking a pause.

This program was later used by a UNIX-script which for the lengths of sequences 2,3,6 sends the text to the program and later sorts the output with respect to frequency and filters out the ten most occurring sequences.

Analysis
A primary selection in order to obtain the target words was made on the basis of the relative frequency within each group. The selected target words were manually replaced in the input of the program by a symbol to signify its new status as a target word. Then the program ran from the beginning with the updated material. A second selection was made in order to obtain additional target words. However, a third selection failed when necessary structure seemed to be lacking to enable any further target words to be found.

Results
The results presented in figures 1a-1b show the ten most frequent sequences with symbols for groups with four and six symbols. The figures are based on the Portuguese material. The results based on the updated material containing already stored words are presented in figure 2a and 2b. Figure 3 show the ten most frequent sequences with symbols for groups with three symbols. This figure is based on the Svensiska material. In figure 1a one can observe the relative frequent occurrences of the symbol sequences $niko$, $onik$, $iko^*$, $eto^*$ and $beto$ in respect to the other symbol sequences. The symbol /*/ is marking a pause, i.e. an utterance boundary.

![Figure 1a. The ten most frequent sequences with four adjacent symbols.](image1.png)

![Figure 1b. The ten most frequent sequences with six adjacent symbols.](image2.png)
Next step, illustrated in figure 1b shows how an additional increase of symbols results in the sequences niko* and beto* becoming the sequences oniko* and obeto* and that these sequences are the most frequent ones. Figure 2a and 2b show the result from the analysis of the updated material after the sequences oniko* and obeto* were replaced by a symbol marking that the word already had been stored. The results show the emergence of a new target word, the sequence *est in figure 2a and *este in figure 2b. Figure 3 showing the ten most frequent symbol sequences from the Svensiska material for the group with three symbols, illustrates how it in this material are fewer frequent sequences with regard to the Portuguese material. No prominent target words could be observed.

Discussion and Conclusion

The aim of this study was to create and evaluate a model, which can give a description of the possible clues infants may rely on when learning their mother tongue. This was made by searching for regularities in sound sequences in infant directed speech. In contrast to adult directed speech, infant directed speech is characterized by a very modulated melody with many pauses and repetition (Sundberg 1998), which justifies the use of infant directed speech in language acquisition studies. The material used in this study is relatively small but still reflects a likely learning situation seeing that the input of the program bears resemblance to the input received by the infant.

Through calculating the relative frequency of the words in the Portuguese material certain words turned out to be more frequent and more prominent than others. Figure 1b also shows how these words almost always end up in utterance final position, furthermore a position that ends with a pause. When the most frequent sequences oniko and obeto were extracted from the material, the program correctly handled them as being target words. These words were marked in the material and would in a learning situation be handled as stored words. When these words were stored a new search was made on the remaining symbol sequences until a new target word, este, emerged. This word however, ended up in an utterance initial position. The search continued on frequent symbol sequences until the search no longer resulted in any new possible target word candidates. Referring these results to language learning, infants hear and notice how certain sound sequences are repeated, much like the computational model. The results in figure 3 also showed that the Svensiska material not received any direct hits for prominent words. This could be due to the relatively large variation within the material.
This variation has to be carefully balanced as too much variation can be difficult for the infant to process. A carefully balanced variation enables the infant to extract underlying linguistic principles (Lacerda & Sundberg, 2001). But maybe one should not necessarily view this flow of information as a problem the infant must struggle with when learning words (Elman 1999), but rather as an asset. When a variety of sensory input are stored simultaneously (i.e. a word and an object) one can assume the likelihood of the two sensory input eventually coincide. A possible development of this study would be to further evaluate the learning model with authentic material from natural mother-child interaction.

A number of computational- and learning models have been suggested but still a clear understanding of what it is that enables mother tongue speakers to transform a continuous speech signal into a sequence with. Lacerda and Lindblom (1996) imagine the infant already from the beginning relying on automatic learning processes to manage information. As the infant stores sensory input and associates information, i.e. auditive and visual, the learning of words get going and soon has created itself a linguistic foundation. On the basis of that foundation, principles for a more elaborated linguistic structure emerges. This linguistic structure is reinforced by the social interaction with the surrounding world.

Our point of view is that the connection of different sensory input in combination with the relative frequency of the sound sequencies, their position within the phrase and the variation in the speech signal together may be of a crucial importance when it comes to the infants’ ability to discover patterns and words in the speech signal. The modified structure of the infant directed speech may also be of help.

References


