Umeå University, Department of Philosophy and Linguistics
PHONUM 9 (2003), *69-72*
Available online at http://www.ling.umu.se/fonetik2003/

Implicit linguistic structure in connected speech

Francisco Lacerda, Lisa Gustafsson and Nina Svärd

Department of Linguistics, Stockholm University

This paper sketches a model of early emergence of basic linguistic structure using general-purpose similarity measures, without *a priori* linguistic knowledge, to structure natural speech signals. The model attempts to mimic a first language learning situation. Crude auditory representations of the acoustic signal are continuously stored in memory and processed to detect similarities between portions of the representation patterns. The similarity measures are purely auditory and allow only for moderate time warping and frequency shifts, picking up any best matches among the stored patterns and between these and incoming speech. An example of the model performance is presented.

1. Introduction

Traditional attempts to mimic human speech processing tend to use sets of rules based on adult linguistic competence. Obviously, integrating full linguistic competence in speech recognition systems is a very good engineering solution but tends nevertheless to fall short when flexibility becomes a crucial demand. Linguistic descriptions essentially capture the adult's performance while ignoring pragmatic aspects, like the speakers' adaptive ability to novel situations and capacity of integrating multiple sources of information. In this context shifting perspective from the crystallized adult linguistic competence to the human infant's ability to achieve it appears as a very powerful strategy that is worth testing. Infants typically discover the underlying structure of their ambient language within only a few years of interaction with their linguistic environment.

The goal of this paper is to show how the initial steps towards linguistic development can be accounted for by general-purpose processes and interaction with the ambient language. We suggest that multi-sensory memory representations along with the multi-dimensional structure of the interactive environment of the infant are the main determinants of the emerging linguistic structure observed within the first years of human development. From a psychoacoustic perspective there are similarities between the infant's and adult's (Werner, 1992) auditory processing of speech sounds but the differences in the infant's and adult's linguistic experience suggests that their interpretations may be radically different. Whereas the adult's linguistic competence tends to force linguistic interpretations on any signal vaguely resembling speech sounds (as demonstrated by perception of sine-wave speech, phoneme restoration, etc), initially the young infant cannot interpret linguistically even the most clearly uttered speech sounds. Knowing that the infant's experience with the ambient language will eventually lead to linguistic development, the issue is how to account for the transition from the acoustically-dominated perception of speech sounds to their overwhelming linguistic interpretation.

1. A general background of the model

We start out by assuming that the infant has an immature memory and can therefore only process simple sentences and structures (Elman, 1999). Improvement in memory capacities and the acquisition of basic primitive representations (the first language-specific sound units stored in the brain) allow the infant to process more complex sentences. However, the input data is also of great importance when the initial capacity of a system is limited but matures with time. Lack of variation in the information may cause the system to make wrong generalizations while too much noise and variation in the information may slow down the learning process until enough data is gathered to make generalizations. Primates in general seem to organize their brain very much depending on what input they are exposed to during their development. Sur et al. (1999) and von Melchner (2000), for instance, carried out a series of studies investigating brain plasticity in primates. They found that young ferrets, which had their visual pathways rewired to their auditory cortex, and their auditory pathways rewired to their visual cortex, began to process visual stimuli with their auditory cortex and vice versa, suggesting that the brain is not as task-specific as previously assumed. (Elman, 1993) proposes that, learning and development interact in important and non-obvious ways. Similarly Minsky (1985) states that task-specific programmed systems cannot learn from their own experience because they have none. And, in this vein, he questions why we make programs do "grown-up" things before we make them do "childish" things. A system must discover and learn basic thinking in order to reach advanced knowledge.

Our model attempts to mimic early language acquisition, where a plastic and immature memory seems crucial for efficiency (Elman, 1999). This follows the model proposed by Svärd, Nehme and Lacerda (2002) that picks up possible word candidates on the basis of their relative frequency in orthographically transcribed speech strings. In this sense "words" are any recurrent chunk of speech and do not necessarily correspond to established concepts although they may function as embryos of further linguistic structure. Initially the infant is assumed to process the speech signal according to crude acoustic variables, like recognizable acoustic boundaries conveyed by relatively sharp overall intensity transitions, and to focus spontaneously on recurrent patterns in the speech signal (Saffran, Aslin, & Newport, 1996). Further linguistic structure, like basic syntactic relations embedded in the input, may also be inferred from transitional probabilities of suggested word candidates and, once discovered, further used to actively explore new speech input.

2. Implementing the model

To represent the initial steps of the natural language acquisition process, we used infant-directed speech (IDS)[1] as input to the current model. A battery of sentences, produced in this IDS style by a female speaker, was digitally recorded in a sound-treated room.[2]

First a crude auditory representation (cf. Carlson & Granström, 1982) is generated by processing the speech signal with a filter bank consisting of 21 digital 1-Bark band-pass filters covering adjacent 1 Bark bands up to 8 kHz, followed by full-wave rectifiers and integrators (figure 1). The level of the input signal was calibrated in dB_{SPL} (i.e. rel. to 20 μPa,

[1] IDS is characterized by high-pitch, highly modulated F_0, pauses and repetitions (Sundberg, 1998).

[2] The signal was sampled at 44 kHz, 16 bit, and subsequently sampled down to 16 kHz before further processing.

1 kHz) and converted to hearing levels (dBHL) by subtracting the hearing thresholds within each Bark band from the original levels.

The "auditory representations" are stored in a general purpose memory buffer that simply remembers salient enough acoustic patterns (i.e. with overall intensity greater than a pre-established threshold, e.g. 25 dB_A), without distinguishing between speech and non-speech sounds. Incoming signals are continuously represented by paths in the 21-dimensional auditory memory space where similarity between different portions of the incoming signals is reflected by nearly overlapping paths.

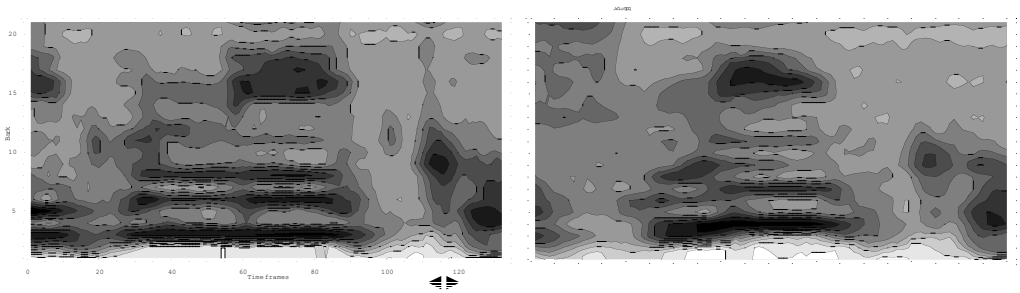

Figure 1. Auditory spectrograms of two instances of [uniku], showing a selected region.

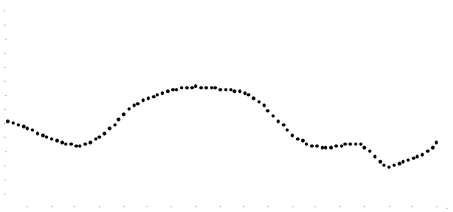

Figure 2. Distance function by sliding the reference pattern marked on figure 1, left, along the auditory representation shown on the right.

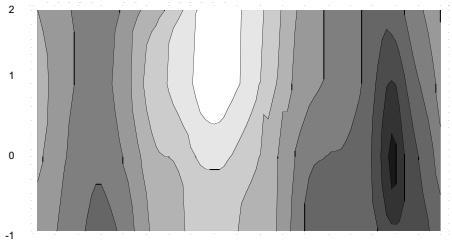

Figure 3. Generalization of the distance function displayed in figure 2, when shifting the original frequency pattern by -1 Bark to 2 Bark.

To illustrate how the model works, we present here the two examples of the sequence [uniku] and after selecting the region marked by the arrows on the left figure 1 we will attempt to find a match, in the right panel. Using a city-block distance metric to measure the similarity between the marked portion on the left panel and any location of the right panel, we get the dB-calibrated distance displayed in figure 2. The contour plot in figure 3 illustrates how a "best match" can be obtained by hovering the reference pattern over the auditory spectrogram of figure 1's right panel. The y-axis shows the relative shift between the frequency scales of the reference and the auditory spectrogram. The x-axis represents the time coordinate (frames) and the darkness of the shaded areas indicates the degree of similarity. Both figures 2 and 3 indicate that the best match with the reference pattern can be found at about frame 80. This is a perfectly reasonable match, detecting the similarity between the explosion bars of the [k]-sounds.

The same procedure can be applied to longer (and potentially more meaningful) chunks of the signal, as will be demonstrated in the presentation of this paper. Increasing the length of the reference sequence tends to lead to more distinct matches, as the probability of spurious matches will decrease as longer chains of frames are compared.

4. Discussion

Our simple example suggests that the nature of the input may be of great importance to developing linguistic representations. Because of the highly repetitive and structured IDS, it may be expected that similarity relations in the input signal enable automatic learning process that may trigger further linguistic development (Lacerda and Lindblom, 1996; Lacerda, 2003). Adding other sensory input to the auditory signal is likely to further enhance the linguistic structure of the infant's ambient language and in this context the infant's immature memory functions "…like a protective veil shielding the infant from stimuli which may either be irrelevant or require prior learning to be interpreted" (Elman, 1993, p95). In further development of our model, we hope to be able to tap on the emergence of syntactic structure. Bates and Goodman (1999), for instance, assume that grammatical structure arises naturally as the child tries to find a way to handle the information flow when accumulating more and more words, a notion compatible with Nowak (2000) approaches to the emergence of grammatical structure.

References

Bates, E., and Goodman, J. C. (1999) On the Emergence of Grammar From the Lexicon. In *The emergence of language* (B. MacWhinney, Ed.). LEA, Publishers, London.

Carlson,R., & Granström,B. (1982). Towards an auditory spectrograph. In *The Representation of Speech in the Peripheral Auditory System* (R. Carlson & B. Granström Eds.), pp. 109-114. Amsterdam: Elsevier Biomedical Press.

Elman, J. L. (1993) Learning and development in neural networks: the importance of starting small. *Cognition*, 48, 71-99.

Elman, J. L. (1999) The Emergence of Language: A Conspiracy Theory. In *The emergence of language* (B.MacWhinney, Ed.). LEA, Publishers, London.

Lacerda, F. (2003) Phonology: An emergent consequence of memory constraints and sensory input, *Reading and Writing: An Interdisciplinary Journal,* 16, 41-59.

Lacerda, F., and Lindblom, B. (1996) Modeling the early stages of language acquisition. In COST-96 (S. Strömqvist, Ed.).

Minsky, M. (1985) Why People Think Computers Can't. In *The Computer Culture*, (Donnelly, Ed.). Ass. Univ. Presses, Cranbury NJ.

Nowak, A. M., Plotkin, B. J., and Jansen, A. A. V. (2000) The evolution of syntactic communication. *Nature*, 404, 495-498.

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

Sundberg, U. (1998) *Mother tongue - Phonetic aspects of infant-directed speech.* PhD Dissertation in Phonetics, Stockholm University.

Sur, M., Angelucci, A., and Sharm, J. (1999) Rewiring Cortex: The role of patterned activity in developmental and plasticity of neocortical circuits. *J. of Neurobiology*, 41, 33-43.

Svärd, N., Nehme, P. and Lacerda, F. (2002) Obtaining linguistic structure in continuous speech*, TMH-QPSR* Vol. 44. Fonetik 2002, Stockholm.

von Melchner, L., Pallas, S. L., and Sur, M. (2000) Visual behavior induced by retinal projections directed to the auditory pathway. *Nature*, 404, 871-875.

Werner,L. (1992) Interpreting Developmental Psychoacoustics. In *Develpmental Psycholinguistics*, (L. Werner & E. Rubel, Eds.), pp. 47-88. Washington: American Psychological Association.