A Computational Model of Preverbal Infant Word Learning

Guillaume Aimetti (G.Aimetti@DCS.Shef.AC.UK)
Speech and Hearing Group, Regent Court, 211 Portobello
Sheffield, S1 4DP UK

Roger K. Moore (R.K.Moore@DCS.Shef.AC.UK)
Speech and Hearing Group, Regent Court, 211 Portobello
Sheffield, S1 4DP UK

Keywords: Early Language Acquisition; Word Learning; Automatic Segmentation; Cross-Modal Association; Dynamic Time Warping.

Introduction

This work investigates a novel computational model of preverbal infant word learning in an attempt to create a more robust speech recognition system. Currently, the state-of-the-art can be extremely accurate when used in its optimal environment. However, when taken out of its comfort zone accuracy significantly deteriorates and does not come anywhere near human speech processing abilities, even for the simplest of tasks. We take inspiration from the ease with which newborns are able to learn words, with no apparent difficulty, and develop into expert communicators of their native language.

In order to learn words, the young language learner must be able to segment speech into useful units and then associate them to visual referents from within their environment (Smith & Yu, 2008). The model described here, the Acoustic DP-ngrams, attempts to solve the word-to-world mapping problem through cross-modal (acoustic & visual) associative learning set within an interactive framework, as illustrated in figure 1 (for a more technical description of the system see (Aimetti, 2009)).

Initial results show that there is significant potential with the current algorithm, as it segments in an unsupervised manner and does not rely on a predefined lexicon or acoustic phone models that constrain current Automatic Speech Recognition (ASR) methods. The learning process concurs with current cognitive views of early language acquisition (Jones, Hughes, & Macken, 2006; Saffran, Aslin, & Newport, 1996; Saffran, Werker, & Werner, 2006; Smith & Yu, 2008), and the key word detection experiments exhibit similar behaviours apparent in developing preverbal infants (Gomez & Gerken, 2000; Kuhl, 2004; Newman, 2008).

The Computational Model

There are two key processes occurring within our learning agent (LA):

1. Automatic Segmentation: Acoustic DP-ngrams is used to automatically segment the speech, directly from the acoustic signal, into important lexical fragments by discovering *similar* repeating patterns. This approach uses a dynamic programming (DP) technique, dynamic time warping (DTW), to accommodate the temporal distortion present in speech. The advantage of this approach is that it uses an accumulative scoring system to measure the quality and length of the discovered fragments. This method is similar to the Segmental DTW algorithm developed to summarise recordings of academic lectures (Park & Glass, 2008).

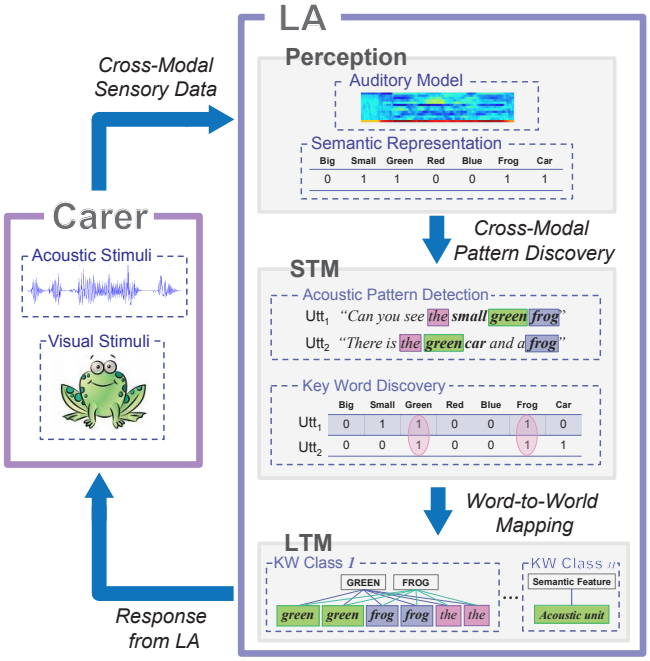

Figure 1: Word-to-World mapping set within an interactive carer-learner framework. LA's internal memory is inspired by current cognitive views (Jones et al., 2006).

2. Word-to-World Mapping: Figure 1 shows the interaction between LA and its parent (carer). During training the carer incrementally feeds LA with cross-modal stimuli; the acoustic stream consists of continuous speech, as sampled data, and the visual stream consists of crisp tags, representing the visual referents within the utterance. Internal representations of the visual referents is achieved through the co-occurring events from both modalities, as suggested by Smith and Yu (2008). Each class is therefore emergent and constantly evolves with the accumulation of exemplar tokens, thus allowing the system to gradually become more robust to the variation present in speech.

Experiments

LA is trained with 480 cross-modal utterances from a single female speaker (F1); each utterance is passed to the system as sampled acoustic data in parallel with the crisp visual tag(s), representing the key word(s) that lie within it. To test the emergence and robustness of internal representations, LA is faced with a recurrent key word detection task throughout development. This is carried out as probe moments which occur every 20 utterances. LA is temporarily frozen and tested on 320 unobserved utterances from the known female speaker (F1) and 320 unobserved utterances from an unknown male speaker (M1). Only the acoustic part of the input is processed and LA must recognise the key word(s), responding with the correct visual referent(s).

Figure 2 displays the key word detection accuracy during the learning period, which is shown as the percentage of correct key word detections over the number of utterances observed. The blue plot with circles shows the F1 probe, the green plot with squares shows the M1 probe and the red discontinuous plot shows the chance level of a correct guess.

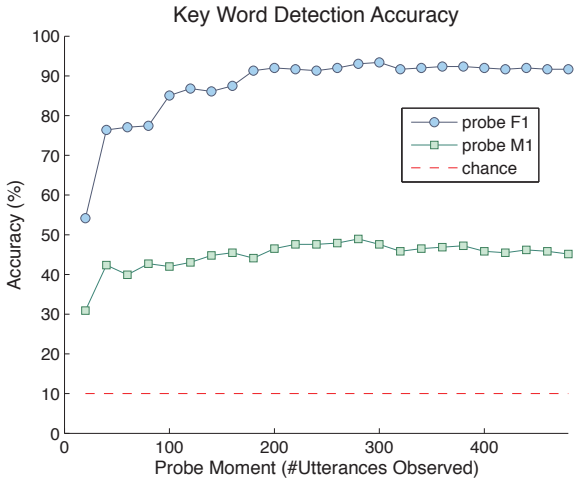

Figure 2: LA's key word detection accuracy throughout development. Probing is carried out every 20 utterances where LA is tested on a known (F1) and unknown (M1) speaker.

Internal representations can be seen to emerge very quickly from the plot in figure 2. After only 20 utterances LA is already able to detect key words well above chance level, achieving 54% for F1 and 31% for M1. Robust representations for F1 develop after 180 utterances, where key word detection accuracy reaches a plateau of $92\%(\pm 1\%)$. However, internal representations for M1 seem to plateau after only 40 utterances and limited to a maximum of 49%.

Discussion & Future Work

This paper introduces a computational model of early word learning abilities in preverbal infants. The algorithm is able to successfully learn words in a cognitively plausible fashion. It is clear to see from the results that LA quickly builds up accurate representations to a familiar speaker F1, but is also still able to generalise above chance level to an unknown speaker M1 across gender with 40%to50% accuracy. This shows that without observing other speakers, the system is not able to build robust internal representations that can reliably generalise across speakers, as suggested by Newman (2008).

One downside to this technique is that it is unable to run on a large data-set as the exemplar tokens being stored in memory are unbound and tend to infinity. Currently, the authors are investigating a method to automatically build prototype representations for the most efficient units within the learners native language (i.e. with Hidden Markov Models). This agrees with current thinking that infants begin learning language attending to too much detail within their native language, and that prototype representations (an average of exemplar units stored in memory) occur with experience from a greater variety of speakers (Kuhl, 2004; Newman, 2008).

Acknowledgments

This research was funded by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org).

References

Aimetti, G. (2009). Modelling early language acquisition skills: Towards a general statistical learning mechanism. In *Proceedings of the student research workshop at eacl 2009* (pp. 1–9). Association for Computational Linguistics.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178–186.

Jones, D. M., Hughes, R. W., & Macken, W. J. (2006). Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory. *Journal of Memory and Language*, *54*(2), 265-281.

Kuhl, P. K. (2004, November). Early language acquisition: cracking the speech code. *Nature*, *5*, 831-843.

Newman, R. S. (2008). The level of detail in infants' word learning. In *Current directions in psychological science* (Vol. 17, p. 229-232). University of Maryland, College Park.

Park, A., & Glass, J. (2008). Unsupervised pattern discovery in speech. In *Trans. alsp* (Vol. 16, p. 186-197).

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *SCIENCE*, *274*, 1926-1928.

Saffran, J. R., Werker, J., & Werner, L. A. (2006). Handbook of child psychology. In D. Kuhn & R. S. Siegler (Eds.), (6th ed., Vols. 2, Cognition, Perception and Language, p. 58-108). New York: Wiley.

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