Discovering Keywords from Cross-Modal Input: Ecological vs. Engineering Methods for Enhancing Acoustic Repetitions

Guillaume Aimetti[1], *Roger K. Moore*[1], *Louis ten Bosch*[2],
Okko Räsänen[3], *Unto K. Laine*[3]

[1]Speech and Hearing Research group, University of Sheffield, UK
[2]Dept. of Language and Speech, Radboud University Nijmegen, NL
[3]Dept. of Signal Processing and Acoustics, Helsinki University of Technology, Finland
G.Aimetti@dcs.shef.ac.uk

Abstract

This paper introduces a computational model that automatically segments acoustic speech data and builds internal representations of keyword classes from cross-modal (acoustic and pseudo-visual) input. Acoustic segmentation is achieved using a novel dynamic time warping technique and the focus of this paper is on recent investigations conducted to enhance the identification of repeating portions of speech. This ongoing research is inspired by current cognitive views of early language acquisition and therefore strives for ecological plausibility in an attempt to build more robust speech recognition systems. Results show that an ad-hoc computationally engineered solution can aid the discovery of repeating acoustic patterns. However, we show that this improvement can be simulated in a more ecologically valid way.

Index Terms: early language acquisition, automatic segmentation, dynamic time warping, speech perception

1. Introduction

There is a growing consensus, within the speech technology field, that gaining a deeper understanding of current cognitive views of early language acquisition and attempting to model these behaviours will help improve the robustness of speech recognition systems. Currently, state-of-the-art systems can achieve very accurate recognition when used in their optimal environment [1]. However, performance drastically deteriorates in the case of mismatching test and training conditions [2].

Not only do newborns tackle the daunting task of language acquisition with ease, but they are also very quick to build robust speech representations of their native language. Current literature within the developmental field suggests that language learning has already started before birth. The fetal auditory system is already functioning by the 25th week of gestation and observations by [3] has shown evidence of fetal memory as early as 32 weeks, through classical conditioning. It is thought that the existence of memory *in utero* is to help recognition and attachment to the mother [4, 3]. Nazzi *et al.* [5] hypothesise that this may be a vital property of early language acquisition. They carried out experiments showing that newborns, as early as five days old, had a preference for the rhythm class of their mother's native language.

Although fetal hearing starts at such an early stage of development, there is a substantial difference between fetal and adult hearing. Sound is strongly filtered by the maternal tissues, amniotic fluid and immature auditory system which act as a low-pass filter. The frequency response of this filter is gradually increased, exposing the infant to ever more complex speech sounds until adult-like hearing is achieved by the end of the first year post birth. However, as Saffran *et al.* [6] note, it is not until early childhood, around 6 years of age, that auditory processing has matured to allow adult-like speech *perception*. Figure 1 displays Lasky & Williams' [7] characterisation of the frequency-specific responsiveness of the fetus compared to newborns and adults. Lasky & Williams assert that this is a speculative depiction of current views of fetal and newborn hearing, as conclusive data does not yet exist. We hypothesise that this gradual unfolding of the auditory environment allows the developing language learner to obtain a coarse grounding for its native language, from which it is then able to bootstrap finer representations with experience.

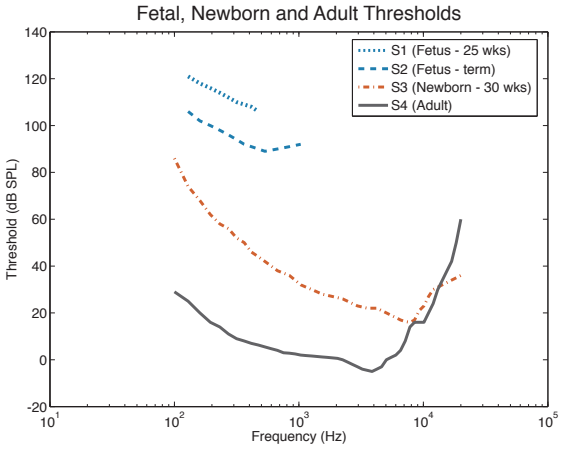

Figure 1: *Fetal, newborn and adult frequency-specific thresholds as speculatively characterised by Lasky and Williams [7] to represent current views of fetal and newborn hearing.*

The work reported here takes inspiration from current anatomical, functional and cognitive views of early speech perception during early language acquisition. We begin by introducing a cognitively motivated computational model of early word learning abilities in preverbal infants. The model derives word meanings by automatically segmenting acoustic speech data into word-like units and mapping them onto discrete visual tags through cross-stuational observations. This model does not begin life with any pre-specified linguistic knowledge,

unlike current Automatic Speech Recognition (ASR) methods which are traditionally trained using a lexicon with phonetic transcriptions in terms of a fixed set of phone-like units. Instead the sharpening of internal representations arises as an emergent property of the cross-modal environment of the system. Segmentation of the acoustic speech signal into word-like units is carried out using a novel dynamic time warping technique. Similar dynamic programming methods have been developed to discover repeating portions of music, with the aid of ad-hoc image processing filtering techniques by reducing the acoustic detail (such as [8, 9]). It is thus interesting to speculate whether such computational solutions can be used to help bootstrap the pattern discovery process in early language learning.

This paper reports the results of an investigation into a comparison of an ecologically inspired approach with an ad-hoc image-processing solution for enhancing repeating portions of speech on a keyword detection task. Results are compared to a baseline system which uses full fidelity acoustic speech data. It is hypothesised that the word error rate should decrease using the implementation of the filtering techniques as the system should be able to find a greater number of longer and more meaningful lexical tokens.

2. A computational model for discovering keywords

Word-like units are learnt in a semi-supervised fashion through the association of co-occurring cross-modal (acoustic and pseudo-visual) events. The learning algorithm (LA) is set within an interactive framework where the caregiver (CA) is able to communicate with LA in a realistic and controlled environment. The following represents a briefly outline of the computational model, and the reader is referred to [10] for a more detailed description. Figure 2 illustrates this interactive environment along with LA's memory architecture, as inspired by current psycholinguistic research [11]:

CA: Provides LA with cross-modal stimuli, which consists of an utterance of sampled acoustic data presented in parallel with an abstract visual tag. This tag is distinct and represents a higher-level perception of the stimulus object within LA's environment. For example, the utterance '*Look at the round **ball***' is associated with the tag ***ball***. It is also important to note that this tag does not give any positional or linguistic information about the keyword within the utterance.

Perception: Converts the acoustic signal into a representation similar to the human auditory system, using Mel-frequency cepstral coefficients as used for conventional ASR (12 MFCC's, energy, delta and delta-delta features).

STM/Working Memory: Stores a limited number of the past n utterances in a short term memory (STM) for carrying out cross-modal associations ($n = 10$ for the reported experiments).

LTM: The cross-modal associations are stored in long term memory (LTM). Acoustic units are mapped to visual tags, allowing LA to build internal representations of its environment.

VLTM: All past utterances are stored in the very long term memory (VLTM). Future work will include additional *sleeping* processes, allowing LA to re-organise internal representations.

2.1. Automatic segmentation

Automatic segmentation is carried out with the Acoustic DP-ngram algorithm [12]. This method uses a popular dynamic programming (DP) technique - dynamic time warping (DTW) - in order to accommodate temporal distortion present in the acoustic speech signal (similar approaches include [13, 14]). Through an accumulative scoring mechanism, this method is able to detect similar portions of speech that commonly re-occur within utterances (such as phones, words and sentences) whilst taking into account noise, speech rate and pronunciation variation. The discovered sub-sequence portions are termed *local alignments* and, once mapped to visual tags in memory, they are termed *lexical tokens*. An additional property of the accumulative quality score is that longer, more meaningful local alignments produce a higher final quality score, thus allowing the system to list lexical units in order of importance. The three steps of the segmentation process are outlined below.

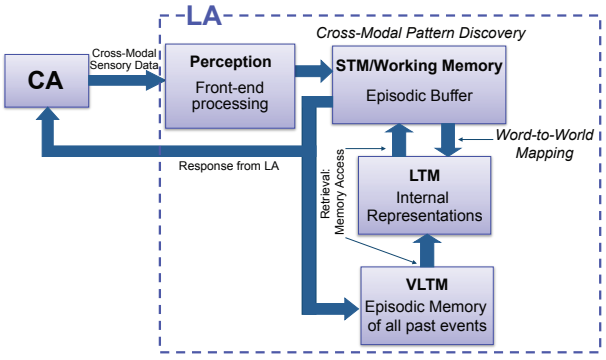

Figure 2: *Diagram of the model framework. The memory architecture is based on current views of memory [11].*

Step 1: Two utterances (utt_1 and utt_2) are fed to STM as two sets of MFCC feature vectors (A, B) where the Euclidean Squared Distance between each pair of frames $d(v_i, v_j)$ for each coefficient c ($n = 39$) is calculated using

$$d(v_i, v_j) = \sum_{c=1}^{n} (A_{ip} - B_{jp})^2 \tag{1}$$

to give the local-match distance matrix D.

Step 2: D is then used to calculate the accumulative quality scores for successive frame steps within A and B using the recurrence defined in Eq. 2 to give the global quality score matrix Q. Higher local quality scores $q_{i,j}$ are achieved by the accumulation of successive local-matches, therefore the score for a local-match must be positive, and scores for both insertions and deletions must be negative in order to penalise temporal distortion (Eq. 3).

$$q_{i,j} = max \begin{cases} q_{i-1,j-1} + (s(a_i, b_j) \cdot d(v_i, v_j)), \\ q_{i,j-1} + (s(\phi, b_j) \cdot |d(v_i, v_{-j}) - 1| \cdot q_{i,j-1}), \\ q_{i-1,j} + (s(a_i, \phi) \cdot |d(v_{-i}, v_j) - 1| \cdot q_{i-1,j}), \\ 0 \end{cases} \tag{2}$$

where,

$$\begin{array}{lll} s(a_i, b_j) & = +1 & \text{(local-match score)} \\ s(\phi, b_j) & = -1 & \text{(insertion score)} \\ s(a_i, \phi) & = -1 & \text{(deletion score)} \\ q_{i,j} & & \text{(local quality score)} \end{array} \tag{3}$$

Backtracking pointers p are maintained at each step of the recursion

$$p_{i,j} = \begin{cases} (i-1, j-1), & \text{(local-match)} \\ (i, j-1), & \text{(insertion)} \\ (i-1, j), & \text{(deletion)} \\ (0, 0) & \text{(initial pointer)} \end{cases} \tag{4}$$

Step 3: Finally, the optimal local alignment is discovered within Q by backtracking from the highest quality score $max(q_{i,j})$ until $q_{i,j}$ equals 0. Multiple local alignments are discovered by repeating this process while $max(q_{i,j})$ is greater than the quality threshold (q_{thresh}).

2.2. Word-to-world mapping

The mapping of local alignments to visual tags is carried out through cross-modal association, as suggested by [15]. The Acoustic DP-ngrams discovers local alignments by comparing two different utterances (utt_1 and utt_2). Therefore, all local alignments can be associated to two different tags (t_1 from utt_1 and t_2 from utt_2). This allows LA to infer that any local alignment with the same t_1 and t_2 can be hypothesised to be keyword t, and if LA does not yet have an internal representation for t then a new keyword class is created for it in LTM. With increasing exposure to incoming data, each keyword class becomes more robust as more lexical tokens with greater variation are obtained.

3. Enhancing repeating acoustic patterns

The basic acoustic DP-ngram algorithm described above performs pattern discovery directly on a full fidelity acoustic signal. This is equivalent to a horizontal frequency response lying at 0Db SPL in figure 1. This configuration is referred to as the 'baseline' system. However, full fidelity frequency resolution can be too fine for the system to detect repeating portions of speech which have a large amount of variation. Therefore, two techniques were investigated for smoothing the variation: an engineering solution using a diagonal filter, and an ecologically-inspired approach using a model of an infant's developing auditory system.

3.1. Diagonal filter

An image processing filtering technique commonly used for smoothing and sharpening two-dimensional images was implemented to highlight repeating acoustic patterns. Areas of high correlation within D are expected to occur along the diagonals, revealing repeating portions of speech from the two utterances being compared. An appropriate filter H is thus specified as

$$H = \begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 1 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 1 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 1 \end{pmatrix} \quad (5)$$

which is optimised for highlighting phone-like repetitions of 75ms and allows for a single insertion/deletion, which is penalised by a 0.5 weighting. The two-dimensional convolution (C) of H and D is specified as

$$c[m,n] = d[m,n] \otimes h[m,n] \quad (6)$$

However, due to the large amount of variation present in speech, there is significant deviation from a perfectly linear diagonal correlation.

3.2. Developing auditory system

A more ecologically valid method for highlighting repeating portions of the acoustic speech data is to model the developing auditory system as characterised by [7] in figure 1. The increasing frequency-specific threshold for each developmental stage (**S1** - Fetus 25 wks, **S2** - Fetus term, **S3** - Newborn 30 wks and **S4** - Adult) is modelled with four 50-channel gammatone filterbanks (g_s). A standard off-the-shelf implementation of the Patterson-Holdsworth auditory filter [16] was used to create the filterbanks by modifying the amplitude gains at different frequencies. It is important to note that the amplitude gains of the frequency-specific thresholds from figure 1 have been shifted so that the minimum point of the filters lie on 0dB SPL. The gammatone filter is described by the equation

$$g_s(t) = a_s t^{n-1} cos(2\pi f t + \phi) e^{-2\pi b t} \quad (7)$$

where a_s is the frequency-specific amplitude for each of the four stages of development, f is the frequency, ϕ is the phase carrier, n is the filter's order (4), b is the filter's bandwidth, and t is time.

4. Data and results

The data used for training consisted of 300 single speaker utterances, each of which contained one of 10 different keywords that LA had to learn (each keyword occurred 20 times). The accuracy of LA's internal keyword representations was measured throughout development with a keyword detection task. LA was only presented with the acoustic part of the utterance and had to recognise the keyword within it, replying with the corresponding visual tag. This was carried out as a probing moment, where LAs internal state was temporarily frozen and tested on 100 unobserved utterances. Probing moments occured more frequently during the early stages of development (every five utterances), allowing the emergence of internal representations to be analysed in finer resolution.

It was hypothesised that the addition of smoothing techniques would aid the automatic segmentation discover a greater number of longer, more meaningful local alignments. Therefore, internal representations were expected to emerge at a faster rate and handle greater variation in the speech due to the larger repository of exemplar tokens.

Table 1 shows the total number of exemplar tokens stored in memory for all internal classes for each LA setting during different stages of development. It is clear that the diagonal filter has accrued the most exemplar tokens, followed by the auditory filters S1 and S2. Each of these configurations discovered more patterns than the baseline system. It is interesting to see auditory filters S3 and S4 hinder the segmentation process, both discovering fewer local alignments than the baseline, S3 yielding the least. This is because S3 is optimised for frequencies $\sim$ 8kHz, therefore processing speech with the greatest variation.

Figure 3 displays the word error rate (WER) as a function

Table 1: *Total number of exemplar tokens stored in LTM for different LA settings.*

	Probe Moment					
	50	100	150	200	250	300
Base	58	134	196	264	334	418
DiagFilt	290	696	1056	1378	1722	2072
S1	144	324	516	698	870	1090
S2	65	144	200	292	370	458
S3	42	98	136	196	250	294
S4	49	126	180	246	316	372

of utterances observed, for the different filtering techniques implemented, against the current baseline. The plots show that both the diagonal filter and the first stage of the auditory system (S1) decrease the WER from the baseline during the early stages of development (probe moment $<= 100$). During this period, both the diagonal filter and S1 achieve an average decrease in WER of 9%. However, stages S2, S3 and S4 are much slower to develop reliable internal representations; S2 and S4 producing an average increase in WER, over the baseline, of 1% and 3% respectively for probe moments $>= 100$. S3 performs the worst, with an average increase in WER of 10% over the baseline for the same period. The probability of a correct guess is 90%, and is plotted as the discontinuous plot.

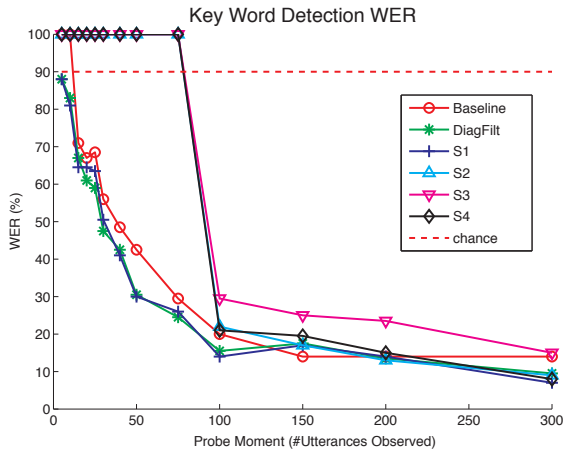

Figure 3: *Comparing keyword WER for the four different stages of the developing auditory system, from fetal to adult-like hearing.*

5. Conclusions

This paper has introduced a novel computational model of early language acquisition, that is able to build accurate internal representations of word-like units with no pre-defined lexical or phonetic information. The experiments show that post-processing image filtering techniques implemented to *force* the appearance of acoustic repetitions can be achieved in a more ecologically valid way. The keyword detection results show that smoothing the acoustic pattern helps the model significantly during the *early* stages of development. However, there seems to be a point at which a larger repository of exemplar tokens, with greater phonetic variation, begins to impede the learners recognition abilities. The advantage of using a frequency-bandwidth limited auditory filter is that the underlying structure of the speech signal is not being distorted in an arbitrary manner (there is always the possibility that the diagonal filter may delete important information from the speech).

A potential disadvantage of the current model is that it is limited to exemplar based recognition. This means that the system is not able to run on a large data-set, as the number of lexical units can increase indefinitely. Therefore, in order to create an efficient and robust speech recognition system, it seem necessary to employ more than just exemplar-based recognition. Current developmental theories suggest that although infants begin life using exemplar representations, they then swiftly adapt to more prototypic units (i.e. an average of the exemplars in memory) [17, 18]. Work is currently being undertaken to try and model this behaviour by first discovering the most efficient units of the native language and then creating appropraite statistical models (e.g. hidden Markov models).

6. Acknowledgement

This research was funded by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org). We are also very grateful for the helpful suggestions and discussions from Tim Kempton regarding the diagonal filter.

7. References

[1] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*. Bristol, PA, USA: Taylor & Francis, Inc., 2002.

[2] J.-C. Junqua and G. van Noord, Eds., *Robustness in language and speech technology*, ser. Text, Speech and Language Technology. Kluwer Academic Publishers, Dordrecht, 2001.

[3] P. G. Hepper, "Fetal memory: Does it exist? what does it do?" *ACTA Paediatrica Supplement*, vol. 416, pp. 16–20, 1996.

[4] A. J. DeCasper and M. J. Spence, "Prenatal maternal speech influences newborns' perception of speech sounds," *Infant Behavior and Development*, vol. 9, pp. 133–150, 1986.

[5] T. Nazzi, J. Bertoncini, and J. Mehler, "Language discrimination by newborns: Toward an understanding of the role of rhythm," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 24, no. 3, pp. 756–766, 1998.

[6] J. R. Saffran, J. Werker, and L. A. Werner, *Handbook of Child Psychology*, 6th ed. New York: Wiley, 2006, vol. 2, Cognition, Perception and Language, ch. The infants auditory world: Hearing, speech and the beginnings of language, pp. 58–108.

[7] R. E. Lasky and A. L. Williams, "The development of the auditory system from conception to term," in *NeoReviews*, vol. 6, no. 3, March 2005, pp. 141–152.

[8] M. Goto, "A chorus-section detecting method for musical audio-signals," in *IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 5, 2003, pp. 437–440.

[9] Y. Shiu, H. Jeong, and C.-C. J. Kuo, "Similar segment detection for music structure analysis via viterbi algorithm," in *IEEE International Conference on Multimedia and Expo*, 2006, pp. 789–792.

[10] G. Aimetti, "Modelling early language acquisition skills: Towards a general statistical learning mechanism," in *Proceedings of the Student Research Workshop at EACL 2009*. Association for Computational Linguistics, 2009, pp. 1–9.

[11] D. M. Jones, R. W. Hughes, and W. J. Macken, "Perceptual organization masquerading as phonological storage: Further support for a perceptual-gestural view of short-term memory," *Journal of Memory and Language*, vol. 54, no. 2, pp. 265–281, 2006.

[12] P. Nowell and R. K. Moore, "The application of dynamic programming techniques to mon-word based topic spotting," *EuroSpeech '95*, pp. 1355–1358, 1995.

[13] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery," in *INTERSPEECH 2007*, 2007.

[14] A. Park and J. Glass, "Unsupervised pattern discovery in speech," in *Trans. ALSP*, vol. 16, no. 1, 2008, pp. 186–197.

[15] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, pp. 1558–1568, 2008.

[16] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Inc, Tech. Rep. 35, 1993.

[17] R. S. Newman, "The level of detail in infants' word learning," in *Current Directions in Psychological Science*, vol. 17, no. 3, University of Maryland, College Park, 2008, pp. 229–232.

[18] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature*, vol. 5, pp. 831–843, November 2004.