Discovering an Optimal Set of Minimally Contrasting Acoustic Speech Units: A Point of Focus for Whole-Word Pattern Matching

Guillaume Aimetti[1], *Roger K. Moore*[1], *L. ten Bosch*[2]

[1]Speech and Hearing Research Group, University of Sheffield, UK
[2]Department of Linguistics, Radboud Univ., Nijmegen, NL

G.Aimetti@dcs.shef.ac.uk

Abstract

This paper presents a computational model that can automatically learn words, made up from emergent sub-word units, with no prior linguistic knowledge. This research is inspired by current cognitive theories of human speech perception, and therefore strives for ecological plausibility with the desire to build more robust speech recognition technology. Firstly, the particulate structure of the raw acoustic speech signal is derived through a novel acoustic segmentation process, the 'acoustic DP-ngram algorithm'. Then, using a cross-modal association learning mechanism, word models are derived as a sequence of the segmented units. An efficient set of sub-word units emerge as a result of a general purpose lossy compression mechanism and the algorithms sensitivity to discriminate acoustic differences. The results show that the system can automatically derive robust word representations and dynamically build reusable sub-word acoustic units with no pre-defined language-specific rules.

Index Terms: speech perception, segmentation, classification

1. Introduction

This paper reports the work being carried out to automatically discover an efficient set of re-usable acoustic speech units that could be used in speech technology, both for recognition and synthesis systems. It has been argued by [1] whether phonemes are the most suitable sub-word unit, and that a finer grained representation may be required. Another argument has been put forward by [2] stating that a fundamentally different approach, to the current state-of-the-art automatic speech recognition (ASR), is required if it is to achieve robust performance anywhere near human speech perception. Therefore, we propose a system that is able to build word models and derive re-usable acoustic units, through cognitively motivated general purpose learning mechanisms, that can be used for word recognition. This work is inspired by [3] who states that certain aspects of language acquisition occur as a result of 'Cognitive efficiency' through simple compression algorithms (Miller's notion of chunking [4]). For an infant to learn language, they must successfully discover, store and recognise repeating patterns from their acoustic surrounding. In order for them to do this *efficiently* the infant must also possess the ability to dynamically optimise these patterns to their constantly changing environment.

An efficient system tries to find the smallest number of units to explain the input by generalising past experience. We hypothesise that phonemic contrasting units *emerge* as a property of an efficient system endowed with the ability to discriminate meaning acoustically. However, the system will not be limited to phonemic units uniquely, but will have the opportunity to derive and employ smaller and larger units, depending on what is more efficient. The Acoustic DP-ngram algorithm [5] is used to discover repeating patterns within the speech signal which is a modification of the algorithm developed by [6] for finding repeated sub-alignments of gene sequences. Through the use of dynamic programming (DP), the algorithm is able to discover and recognise similar, but not necessarily identical, acoustic patterns.

It does not seem cognitively plausible that a language learning system is storing whole words uniquely. Remembering multiple exemplars for each word a human hears in their entire life would require an enormous memory capacity and an extremely efficient recognition mechanism. This process seems very wasteful as words are made up of a finite number of repeating acoustic patterns, albeit with lots of co-articulation. This raises the important question of **what are the fundamental units of speech used by humans?**

The hypothesis is that phonemic contrasting units *emerge* as a property of a system pre-wired with two essential needs:

1. The need to compress sensory information as much as possible to maximise efficiency
2. The need to discriminate between sensory events of different meaning to maximise understanding

The algorithm presented in this paper demonstrates an example of a word-learning mechanism that is able to discover reusable sub-word units. The experiment presented here has been inspired by previous work showing that the discrimination between two similar sounding words can be greatly improved by using a network-type data structure [7]. However, the advantage of the acoustic DP-ngrams is that it is a more general purpose pattern discovery algorithm that learns in an incremental online manner.

This investigation looks at the discrimination ability of a traditional dynamic programming whole-word pattern matching technique compared to acoustic DP-ngrams for the pair of similar sounding words "stalagmite" and "stalactite". Both algorithms use the same training set from a single male speaker and are tested with unobserved utterances from two different speakers, one male and one female.

The paper is organised as follows. Discovering the particulate structure of speech with acoustic DP-ngrams is described in section 2. The cross-modal word learning process and the emergence of a set of re-usable sub-word units are described in section 3. Experiments are presented in section 4, followed by the results in section 5. Finally, the paper is concluded along with further work in section 6.

2. Discovering the particulate structure of speech

2.1. The data

Three different native English speakers were recorded for this investigation: A_m, B_f and C_m (subscript denotes gender). Each speaker uttered 120 examples of each of the pair "stalagmite" and "stalactite", which was recorded at a sampling frequency of 16kHz and then manually segmented. Ten pairs were chosen at random from speaker C_m as the training set, while A_m and B_f were used as the speaker-independent test set, consisting of a total of 480 word examples. The pair of words "stalagmite" and "stalactite" were chosen as they are two similar sounding words that only differ in the middle of the word. Both words differ over two consecutive phones /gm/ and /kt/.

Front-end analysis was carried out to obtain a frame-by-frame 10 ms spaced 13-dimensional vector representation (V), 12 mel-frequency cepstral coefficients (MFCC's) along with energy, using common parameters for speech processing.

2.2. Pattern discovery

The acoustic DP-ngram algorithm detects repeating portions of the acoustic speech signal, through a dynamic programming (DP) technique (cf. [5]). Traditionally, DP was used for whole-word recognition by finding the shortest distance between an acoustic input and a set of templates. However, the current method uses an accumulative quality scoring mechanism in order to reveal repeating sub-portions of two acoustic signals, termed *local alignments*. The main steps of the process are:

Step 1: Each utterance utt_i is input for the learner as acoustic feature vectors and is compared against exemplars stored in memory ($E = \{e_1, \ldots, e_k\}$). For each pair of utterances ($\{V_i\}_{i=1}^{m}, \{V_j\}_{j=1}^{n}$), the Euclidean squared distance between each pair of frames $d_{i,j}$ is calculated using

$$d_{i,j} = \sum_{p=0}^{q} (V_{ip} - V_{jp})^2 \tag{1}$$

where $q = 13$ MFFC dimensions in each frame. When the distance in each cell has been calculated, $d_{i,j}$ is scaled to values between [0,1] to allow for weighted penalty scores.

Step 2: The recurrence defined by Eq. (2) produces local quality scores ($q_{i,j}$) for each cell in $d_{i,j}$. In order to maximise on local alignment length, local-match scores ($s_{i,j}$) must be positive, and both insertion ($s_{\phi,j}$) and deletion ($s_{i,\phi}$) scores must be negative. The values of $s_{\phi,j} = +1$, $s_{\phi,j} = -1$ and $s_{i,\phi} = -1$ were used in this paper. Backtracking pointers (b_i, b_j) are maintained at each step of the recursion.

$$q_{i,j} = max \begin{cases} q_{i-1,j-1} + (s_{i,j} \cdot d_{i,j}), \\ q_{i,j-1} + (s_{\phi,j} \cdot (1 - d_{i,j-1}) \cdot q_{i,j-1}), \\ q_{i-1,j} + (s_{i,\phi} \cdot (1 - d_{i-1,j}) \cdot q_{i-1,j}), \\ 0 \end{cases} \tag{2}$$

Step 3: The optimal local alignment is discovered within q by backtracking from the highest quality score $max(q)$ until $q_{b_i,b_j} = 0$. At each backtracking step q_{b_i,b_j} is set to 0, as well as neighbouring cells that follow in time with a lower quality score, i.e. $q_{b_{i+1},b_j} = 0$ if $q_{b_{i+1},b_j} < q_{b_i,b_j}$. This process is essential for eliminating the extraction of many alignments with slight variations. Multiple local alignments are discovered by repeating this process while $max(q)$ is greater than the quality threshold (q_{thresh}) to output the set $X = \{x_1, \ldots, x_n\}$. Figure 1 displays q, the darker areas show correlation between the two sequences and longer stretches accrue a higher final quality score. Extracted local alignments are shown as the black lines, a thicker line represents a higher final quality score.

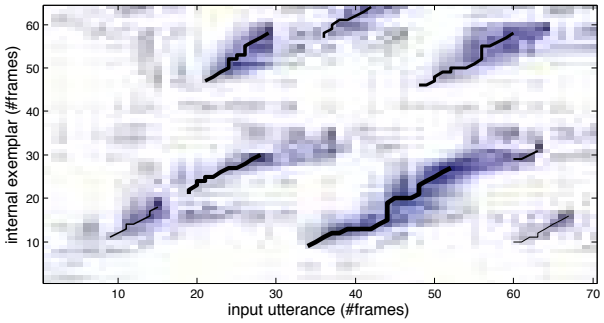


Figure 1: *Quality score matrix calculated using the acoustic DP-ngram algorithm. Darker areas show a higher quality score and the thick black lines show the retrieved local alignments. The thicker the line, the higher the final quality score.*

Once all of the local alignments (X) have been retrieved, they are clustered with the elements of E based on acoustic similarity through hierarchical agglomerative clustering. This clustering process is used as it does not require initialising the number of clusters *a priori*. Acoustic similarity between two elements of E is the minimum-cost alignment between the two sequences, derived using DP, which is then normalised with frame length to give a distance value h between 0 and 1 (where 0 is similar and 1 is different). The algorithm begins by first initialising each element of X and E as separate clusters $\{C_1, \ldots, C_k\}$ of size 1, then the two clusters C_i and C_j with the shortest distance are merged together to create $k-1$ clusters. Complete linkage clustering is used, thus the distance between a pair of clusters is computed as the distance measure between the two furthest elements from the two clusters, as defined by

$$H(C_i, C_j) = \max_{c_a \in C_i, c_b \in C_j} \big[h(c_a, c_b)\big]. \tag{3}$$

This process continues until the distance threshold tau is reached ($0 \leq tau \leq 1$). Each cluster is represented by the cluster centroid, which is the alignment with the shortest distance from all the others within the same cluster. This is now the new set of internal exemplars (E), which dynamically evolves after every utterance.

Varying tau greatly affects the behaviour of this system. If tau is set too high, the system cannot differentiate acoustic sounds and everything is clustered as the same class. If tau is set too low, the system will class all acoustic sounds as different and create many redundant classes. Thus, tau not only controls the systems ability to differentiate similar speech patterns, but also dictates the efficiency of the system, as the number of classes in E is desired to be as small as possible.

2.3. Pattern recognition

Recognition is carried out by finding the optimal path through the input utterance using E. The quality matrix, calculated in the discovery stage, from each internal episodic exemplar (Q_E) is used in order to keep the recognition and discovery process unified and reduce additional parameters. The optimal path through Q_E is derived using DP, however, instead of finding the

minimum cost path, we search for the maximum accumulative quality score. In order to accumulate the quality score across exemplars, we allow the score at the end of one exemplar to be carried over to the beginning of the next.

Figure 2 shows the optimal path (the thick continuous line) through the input utterance. The x-axis displays the input utterance and the y-axis displays the set E. Exemplar boundaries have been marked out using a dotted line, and it can be seen that exemplar jumps only occur at a boundary. The end frame of the y-axis is an additional *unknown state* which is used when a portion of the input cannot be explained. Sequences of unknown states are then appended to E as potentially useful exemplars for future use, thus allowing the system to handle 'out-of-vocabulary' tokens.

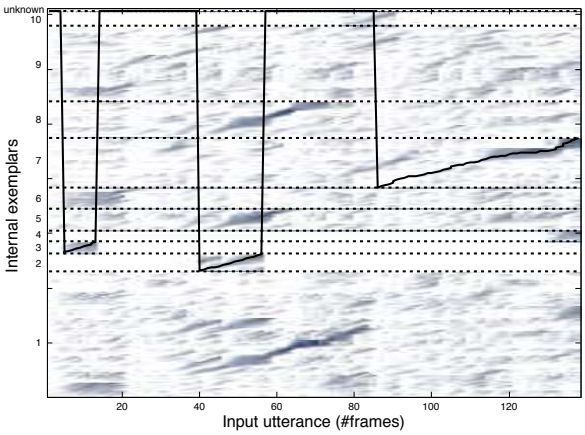


Figure 2: *Optimal path through the input utterance using the quality score matrix from all internal exemplars (E). The x-axis shows the input utterance and the y-axis displays the set E. The end frame on the y-axis is an* unknown state*, which allows the algorithm to store unrecognised portions of speech.*

Miller's notion of chunking states that frequently occuring groups are preferred to less frequent ones, and large chunks are preferred to small chunks [4]. This problem is solved by adding a cost for jumping out of one exemplar and into another. Without this cost the system would use the first exemplar it reaches when at a boundary or in the unknown state. Adding the exemplar jump cost gives the system a preference for longer exemplars as using too many smaller ones will not be optimal. However, if the cost is too high, then the best path will always prefer the unknown state.

3. Word-to-world mapping

Within this experiment there are only two words to be learnt, *stalagmite* and *stalactite*. Each word from the data set is grounded with its corresponding discrete label, which, for the test set will be hidden from the recognition system. However, during training the label is present and allows the system to build word models. The word models are derived by associating the label, of the current training utterance, with the sequence of classes of E that were used to obtain the optimal path during the recognition stage.

As an example, at the start, the system does not have any internal classes to recognise the first training utterance, which is *stalagmite*. But as mentioned earlier in section 2.3, because of the *unknown state*, the algorithm will store the whole training template as its first model. The second training utterance is *stalactite*, from which the pattern discovery process will discover the repeating units [stala] and [ite] and store these as new classes in E. During the recognition process the optimal path would be [stala - *unknown state* - ite], from which the system will now have discovered the class [gm].

4. Experiment

Acoustic DP-ngrams is an exemplar word learning approach that retains a lot of acoustic information, thus efficiency is very important. Varying the normalised distance threshold tau causes the system to derive different sets of sub-word units E. The aim of the experiment was to derive the smallest set of acoustic classes (E) whilst retaining the ability to discriminate meaning between the two similar sounding words "stalagmite" and "stalactite". The different values of tau under investigation were between 0.3 and 0.02 (inclusive) in increments of 0.02. This tau-region was chosen as it appeared critical.

The system was compared against a common whole-word DP pattern matching technique with two performance measures (REF and BASE). The first training example of each word is used as a reference template for the REF baseline. The complete set of twenty training utterances (ten for each word) are used for the BASE baseline and also to train the acoustic DP-ngrams. Each system is then run to recognise the complete test set.

5. Results

Figure 3 displays the mean percentage word error rates (WER) for the complete test set (fig. 3(a)) and both speakers A_m and B_f plotted individually (fig. 3(b)). Also annotated on figure 3(a) is the total number of different classes E derived by the acoustic DP-ngram algorithm for the various settings of tau.

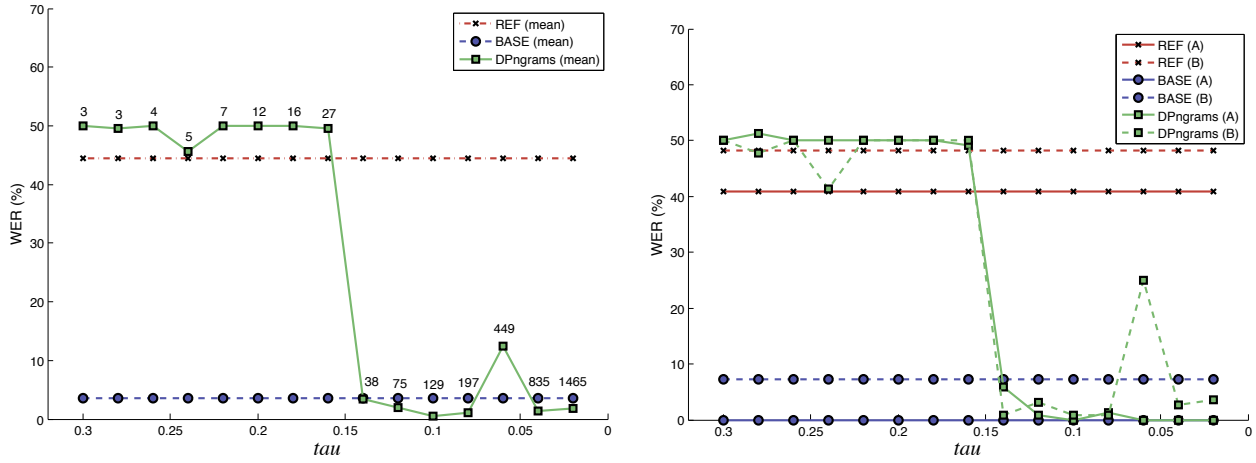
For both plots the square markers display the acoustic DP-ngrams, the circular markers display the BASE baseline and the cross markers display the REF baseline. In figure 3(b) the continuous line displays the WER for the test set A_m and the dotted line displays the WER for the test set B_f.

The mean WER baseline for the complete data set is 44.5% for REF and 3.6% for BASE (fig 3(a)). This shows that the use of a larger training set significantly improves word recognition accuracy (McNemar test [8], $P \ll 0.001$). The mean WER baseline for the individual speakers is presented in figure 3(b), REF achieved 40.8% for A_m and 48.2% for B_f, BASE achieved 0% for A_m and 7.3% for B_f. This shows that with additional training examples we can achieve perfect speaker independent word recognition, of the same gender, and decrease WER for speakers of a different gender.

Acoustic DP-ngrams significantly outperforms the BASE baseline with certain tau settings (McNemar test [8], $tau = 0.1, P \ll 0.001$; $tau = 0.04, 0.02, P \ll 0.01$; $tau = 0.08, P < 0.05$). With tau set to 0.1 the WER is 0.5% for the complete test set. The most efficient number of classes to discriminate the two words and handle acoustic variation is 129.

Whilst the traditional DP template matching technique achieved a decrease in WER across gender with the addition of the training set, the acoustic DP-ngrams achieves a WER of 0.5% for B_f with tau set to 0.1 (fig. 3(b)). Compared to the BASE baseline, this is a significant decrease in WER (McNemar test [8], $tau = 0.1, P \ll 0.001$). This shows that acoustic DP-ngrams is less speaker and gender dependant.

It is also interesting to note that as tau is decreased down to and including 0.08, the WER is consistent for both of the



(a) *Mean WER for the complete test set. The total number of internal classes (E) for the DP-ngrams is labelled.*

(b) *Mean WER for the individual test speakers A_m and B_f.*

Figure 3: *These plots show a) the mean WER for the complete test set and b) the mean WER for the individual test speakers A_m and B_f (240 utterances each). Training was carried out with ten random templates of each word from speaker C_m.*

test speakers. The biggest difference is when tau is decreased below 0.08, where the mean WER remains at 0% for speaker A_m but rises to 25% for speaker B_f when tau is set to 0.06. This shows that increasing the system's ability to differentiate finer acoustic sounds reduces its ability to generalise across speakers.

It can also be seen from the two plots in figure 3 that if tau is too great (i.e. greater than 0.18), then the system is not capable of discriminating the acoustic difference between the two words. In this case the first model that has been built dominates and the system only ever recognises a single word, hence a WER of 50%. As tau decreases, the systems ability to discriminate finer acoustic detail increases, this means that more classes in E are emerging. The increase of the number of different classes is exponential and necessitates a form of pruning. Not only technically, but also conceptually.

6. Discussion and conclusions

This work presents a general statistical learning mechanism for automatically discovering an efficient set of re-usable sub-word units. The system learns words without any *a priori* linguistic knowledge, thus removing the need for pre-specification from a human expert. This system learns in a data-driven, dynamic and incremental manner. To its advantage, it is not constrained to pre-specified lexical units and therefore successfully manages 'out-of-vocabulary' input. The results have shown that acoustic DP-ngrams significantly outperforms the BASE and REF baseline, handling non-uniform speech variation between speakers and gender.

The acoustic DP-ngram process allows the system to automatically build a suitable lexicon for its native language (as well as others), taking into account speech variation. The sub-word units arise as an emergent property of the system interacting with its environment and striving for efficiency without compromising its ability to differentiate meaning. These exemplars could be considered 'phonemic' as they are derived acoustically and semantically. Thus, this approach takes a small step towards providing evidence for an empirical solution for discovering the fundamental units of speech used by humans. Additionally, the automatically derived units could also be used to train current ASR or speech synthesis systems.

The next stage of this research is to record a larger data set with a greater level of speech variation. The data set will include examples of the two words "stalagmite" and "stalactite" recorded, firstly, from read speech from a coherent text and, secondly, from natural speech from a conversation between two speakers on the subject of stalagmites and stalactites. We hypothesise that the recognition ability of acoustic DP-ngrams will be robust against a greater amount of speech variation.

7. Acknowledgements

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

8. References

[1] M. Ostendorf, "Moving away from the 'beads-on-a-string' model of speech," in *Proc. IEEE ASRU Workshop 1999*, 1999.

[2] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," in *Proc. EUROSPEECH 2003, Geneva*, 2003, pp. 2582–2584.

[3] J. G. Wolff, "Language acquisition, data compression and generalization," *Language and Communication*, vol. 2, pp. 57–89, 1982.

[4] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *The Psychological Review*, vol. 63, pp. 81–97, 1956.

[5] 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.

[6] T. Smith and M. Waterman, "Identification of common molecular subsequences," *J. Mol. Bwl*, vol. 147, pp. 195–197, 1981.

[7] R. K. Moore, M. J. Russel, and M. J. Tomlinson, "The discriminative network; a mechanism for focusing recognition in whole-word pattern matching," in *ICASSP 83, Boston*, 1983, pp. 1041–1044.

[8] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. IEEE Conf. on Acoustics, Speech and Sig. Proc. 1989*, 1989, pp. 532–535.