Modelling the emergence of speech sound categories in evolving connectionist systems

J. Taylor[1], N. Kasabov[2] and R. I. Kilgour[2]
[1]Department of Linguistics
[2]Department of Information Science
University of Otago
P.O.Box 56 Dunedin, New Zealand

Abstract *- We report on the clustering of nodes in internally represented acoustic space. Learners of different languages partition perceptual space distinctly. Here, an Evolving Connectionist-Based System (ECOS) is used to model the perceptual space of New Zealand English. Currently, the system evolves in an unsupervised, self-organising manner. The perceptual space can be visualised, and the important features of the input patterns analysed. Additionally, the path of the internal representations can be seen. The results here will be used to develop a supervised system that can be used for speech recognition based on the evolved, internal sub-word units.*

1. Introduction

Competent speakers of a language hear their language, not as a continuously changing stream of sound, but as a succession of discrete, meaning-bearing units. That is, words, or word-like elements. The words themselves are heard, not as unique, globally differentiated patterns of sound variation, but as structured sequences of smaller sound units, which are in themselves meaningless. While the set of words in a language is very large, and potentially open-ended, the number of sound units, or phonemes, is quite small, and relatively stable, even across different accents of the same language. Some languages, such as Māori and Japanese, make do with about twenty phonemes; some languages have well over a hundred. As the languages of the world go, English, with about 45 phonemes, is about average. As every foreign language student knows, languages differ significantly with respect to their phonological organisation --- that is why it is so difficult for a speaker of one language to acquire a native-like accent in a foreign language. Speakers of different languages tend to “hear” the foreign language sounds through the categories of their native language.

Although competent speakers of a language hear, and conceptualise, their language in terms of discrete units (words and phonemes), the acoustic signal bears no signs of discrete segmentation into words or phonemes. Phoneme categories are abstractions some way removed from the raw acoustic data. At the same time, given the language specificity of phonological organisation, it is evident that phoneme categories have to be acquired on the basis of exposure to the input language,

1.1 Perceptual Space

Research by Jusczyk [1], Kuhl [2], and others, has shown that new-born infants are able to discriminate a large number of speech sounds. In fact, well in excess of the number of phonetic contrasts that are exploited in the language an infant will subsequently acquire. This is all the more remarkable, since the infant vocal tract is physically incapable of producing adult-like speech sounds [3]. The ability to discriminate sounds must therefore be based on purely auditory analysis, and cannot be attributed to a feedback loop from articulation (cf. the ‘motor theory’ of perception [4]). By about 6 months, perceptual abilities are beginning to adapt to the environmental language, and the ability to discriminate phonetic contrasts that are not utilised in the environmental language declines. At the same time, and especially in the case of vowels, acoustically different sounds begin to cluster around perceptual prototypes, which correspond to the emerging phoneme categories of the target language [2]. Thus, the ‘perceptual space’ of, for example, the Japanese or Spanish learner becomes increasingly distinct from the perceptual space of the English or Swedish- learner: Japanese, Spanish, English, and Swedish ‘cut up’ the acoustic space differently, with Japanese and Spanish having far fewer vowel categories than English and Swedish. It would appear that the emergence of phoneme categories is driven not only by acoustic resemblance. Kuhl's research showed that infants are able to filter out speaker-dependent differences, and attend only to the linguistically significant phoneme categories.

1.2 Self-Organisation

A central issue in language acquisition research concerns the richness of the initial state. The dominant view within Linguistics has been that the general architecture of language is innate, the learner only requires minimal exposure to data in order to set the open parameters given by Universal Grammar [5]. Recently, this view has been challenged, with greater emphasis being placed on the role of a learning mechanism which generalises over rich arrays of input data [6,7]. In computational terms, the contrast is between highly supervised systems with a rich in-built structure, and minimally supervised, self-organising systems. Research on the latter is still in its infancy, and has been largely restricted to modelling circumscribed aspects of morphology and syntax, most notably, the acquisition of regular and irregular verb morphology [8].

The experiments reported here are part of a larger project, which attempts to model phonological acquisition under conditions of minimal supervision. The project aims to test the hypothesis that language learning takes place through incremental, on-line self-organisation of natural language input. The initial state is an unstructured, multi-dimensional internal acoustic space. Input words are represented as pathways of nodes through the multidimensional space. Repeated tokens of a word type are presented by a band of pathways, while different word types are presented as differentiated pathways. We hypothesise that the trajectories representing different word types may partially overlap, to the extent that different word types share common phonemic constituents.

In this paper, we report on the clustering of nodes in internally represented acoustic space. The emerging nodes correspond to emerging sound types, but may not necessarily correspond to the phoneme categories. Research on the internal representation of word types, and on the emergence of sound categories that may be comparable to the phonemes, is in progress.

2. Evolving Neural Systems

2.1 The ECOS paradigm

ECOS are systems that evolve in time through interaction with the environment; That is, an ECOS adjusts its structure with a reference to the environment [9-11]. ECOS are multi-level, multi-modular structures where many modules have inter-and intra-connections. The evolving connectionist system does not have a clear multi-layer structure. It has a modular open structure. The functioning of the ECOS is based on the following general principles [9-11]:

(1) fast learning from a large amount of data, e.g. through one-pass training;
(2) adaptation in an on-line mode where new data is incrementally accommodated;

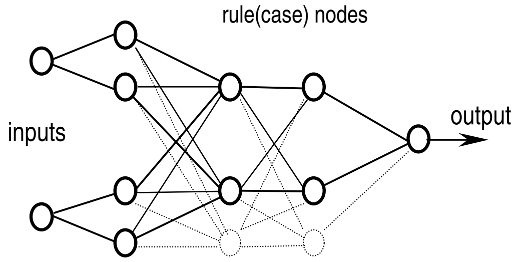

Figure 1: Structure of ECOS system

(3) ‘open’ structure where new features (relevant to the task) can be introduced at any stage of the system's operation, e.g., the system creates “on the fly” new inputs, new outputs, new modules and connections;
(4) memorising data exemplars for a further refinement, or for information retrieval;
(5) learn and improve through active interaction with other IS and with the environment in a multi-modular, hierarchical fashion;
(6) adequately represent space and time in their different scales; have parameters that represent short-term and long-term memory, age, forgetting, etc.;
(7) deal with knowledge in its different forms (e.g., rules; probabilities); analyse itself in terms of behaviour, global error and success; “explain” what the system has learned and what it “knows” about the problem it is trained to solve; make decisions for a further improvement.

2.2. Evolving fuzzy neural networks for supervised and unsupervised learning

EFuNNs are introduced in [9-11]. EFuNNs are models for evolving supervised learning from data that have five-layer structure where nodes and connections are created/connected as data examples are presented (see Figure 1). An optional short-term memory layer can be used through a feedback connection from the rule (or 'case') node layer. The third layer of neurons (rule nodes) in EFuNN evolves through either supervised (EFuNNsu) or unsupervised (EFuNNun) learning. In the experiments presented in this paper we use EfuNNun.

3. Experiments

3.1 Method

To create the clustered model for New Zealand English, several speakers from the Otago Speech Corpus [12] were selected to train the system. Here, 18 speakers (9 Male, 9 Female) spoke 128 words each three times. Thus, approximately 6912 utterances were available for training.

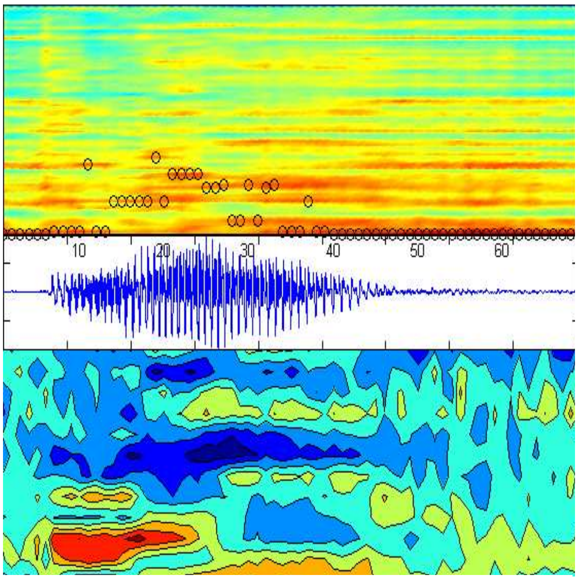

Figure 2: Representation of a spoken word: ‘zero’

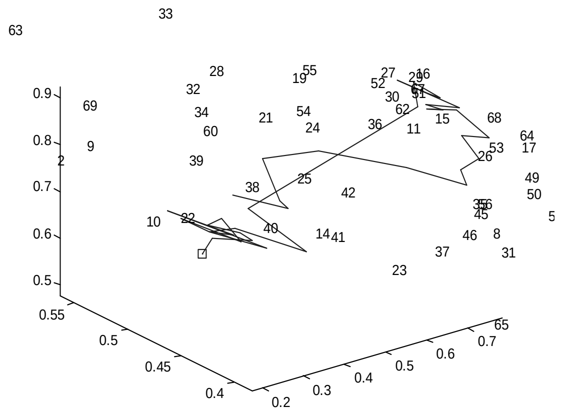

Figure 3: Trajectory of a spoken word: ‘sue’

During the training, a word example was chosen at random from the available words. The waveform underwent a Mel-scale cepstrum (MSC) transformation to extract 12 frequency coefficients, plus the log energy, from segments of approximately 23.2ms of data. These segments were overlapped by 50%. Additionally, the delta and delta-delta values of the MSC coefficients and log energy were extracted, for an input vector of dimensionality 39.

3.2 Results

The system was trained until the number of rules was constant for over 100 epochs. A total of 12000 epochs were performed. The parameters were set to *Sthr* of 0.85. The aggregation threshold was allowed to change, with a target number of rule nodes of 100. The other parameters were as their default values.

Figure 2 shows three representations of a spoken work from the corpus. Firstly, the word is viewed as a

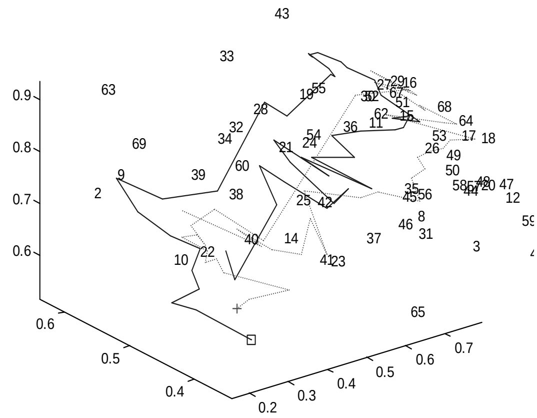

Figure 4: Two utterances of the word ‘sue’

waveform (Figure 2, middle). This is the raw signal as amplitude over time.

The second view is the MSC space view. Here, the 12 frequency components are shown (Figure 2, bottom). This approximates a spectrogram.

The third view (Figure 2, top) shows the activation of each of the rule nodes over time. In this system, 70 rule nodes were created. Darker areas represent a high activation. Additionally, the winning rules are shown as circles. Numerically, these are: 1 1 1 1 1 1 2 2 2 2 22 2 2 11 11 11 11 11 24 11 19 19 19 19 15 15 16 5 5 16 5 15 16 2 2 2 11 2 2 1 1 1...

Some further testing showed that recognition of words depended on not only the winning rule node, but also the path of the recognition. Additionally, an *n*-best selection of rule nodes may increase discrimination.

3.3 Trajectory plots

The trajectory plots, shown in Figures a, b, and c, are in three dimensions of the 39 possible. Here, the first and seventh MSC are used for the *x* and *y* co-ordinates. The log energy is represented by the *z*-axis.

A single word, ‘sue’, is shown in Figure 3. The starting point is shown as a square. Several frames represent the hissing sound, which has low log energy. The vowel sound has increased energy, which fades out toward the end of the utterance.

Two additional instances of the same word, spoken by the same speaker, are shown in Figure 4. Here, a similar trajectory can be seen. However, the differences in the trajectories represent the intra-speaker variation.

Inter-word variability can be seen in Figure 5, which shows the ‘sue’ from Figure 2 (dotted line) compared with the same speaker uttering the word ‘nine’. Even in the three-dimensional space shown here, the words are markedly different.

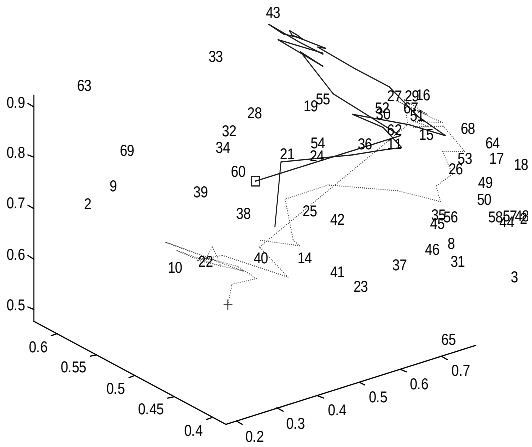

Figure 5: Trajectories of ‘sue’ and ‘nine’

The final trajectory plot (Figure 6) is of two similar words, ‘sue’ (dotted line) and ‘zoo’ (solid line) spoken by the same speaker. Here, there is a large overlap between the words, especially in the latter section, the vowel sound.

4. Future work

The ECOS paradigm is appropriate to modelling emergence of acoustic sound clusters. The next step of the project is to evolve these clusters in a supervised mode of learning with the use of EFuNNsu when words are used as desired outputs for the system to learn. The evolved system will be used as a word recognition system. It will follow the principles for building adaptive speech recognition systems given in [13,14].

Acknowledgements

This work has been funded by a Divisional Research Grant, Humanities, University of Otago, New Zealand.

References

[1] P. Jusczyk, *The Discovery of Spoken Language*, Cambridge, MA: MIT Press, 1997.

[2] P. K. Kuhl, "Speech Perception," in *Introduction to Communication Sciences and Disorders*, F. Minifie, Ed., San Diego, CA: Singular Pub Group, 1994, pp. 77-142.

[3] P. Lieberman, *Uniquely Human: The Evolution of Speech, Thought, and Selfless Behavior*, Cambridge, MA: Harvard University Press, 1991

[4] Liberman, *Speech: A Special Code*, Cambridge, MA: MIT Press, 1996.

[5] N. Chomsky, *The Minimalist Program*, Cambridge, MA: MIT Press, 1995.

[6] M. S. Seidenberg, "Language acquisition and use: Learning and applying probabilistic constraints," *Science*, vol. 275, pp. 1599-1603, 1997.

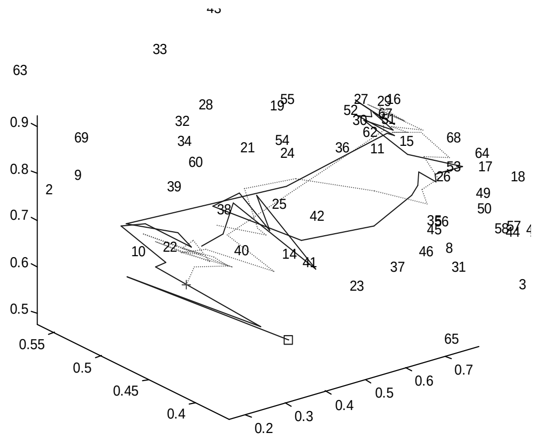

Figure 6: The words ‘sue’ and ‘zoo’

[7] E. Bates and J. Elman, "Learning rediscovered," *Science*, vol. 274, pp 1849-1850, 1996.

[8] K. Plunkett, "Connectionist approaches to language acquisition," in *The Handbook of Child Language*, P. Fletcher and B. MacWhinney, Eds., Oxford: Blackwell, 1995, pp. 36-72.

[9] N. Kasabov, "The ECOS framework and the 'eco' training method for evolving connectionist systems," *Journal of Advanced Computational Intelligence*, vol. 2, no. 6, pp. 195-202, 1998.

[10] N. Kasabov, "Evolving fuzzy neural networks: Theory and applications for on-line adaptive prediction, decision making and control," *Australian Journal of Intelligent Information Processing Systems*, vol. 5 (3), pp. 154-160, 1998.

[11] N. Kasabov, "Evolving connectionist and fuzzy connectionist systems – theory and applications for adaptive, on-line intelligent systems," in *Neuro-Fuzzy Techniques for Intelligent Information Systems*, N. Kasabov and R. Kozma, Eds., Heidelberg: Physica Verlag, 1999, pp. 111-146.

[12] S. Sinclair, and C. Watson, "The Development of the Otago Speech Database," in *Proceedings of ANNES '95*, 1995, pp. 298-301.

[13] N. Kasabov, R. Kilgour and S. Sinclair, "From hybrid adjustable neuro-fuzzy systems to adaptive connectionist-based systems for phoneme and word recognition," *Fuzzy Sets and Systems*, 130 (2), 1999.

[14] N. Kasabov, "A framework for intelligent conscious machines and its application to multilingual speech recognition systems," *Brain-like computing and intelligent information systems*, S. Amari and N. Kasabov, Eds., Singapore: Springer Verlag, 1998.