Modeling the emergence of bilingual acoustic clusters: A preliminary case study

Mark Laws[11], Richard Kilgour [2], Nikola Kasabov[1]

[1] *Knowledge Engineering & Discovery Research Institute, Auckland University of Technology, Auckland, New Zealand.*

[2]*NAVMAN New Zealand Ltd, Auckland, New Zealand.*

Abstract

This paper presents some preliminary results of an original study to model the emergence of bilingual acoustic clusters of both New Zealand English and New Zealand Maori speech. This is performed using true on-line learning in a connectionist architecture. The study represents a joint collaborative analysis, which applies the bilingual data as training examples to a connectionist-based evolving clustering method algorithm. The algorithm returns a structure containing acoustic clusters plotted using visualization techniques that could be used as the foundations for future speech classification systems. The following experiments are based on the notion that approximately 75% of the phonological units in New Zealand English and New Zealand Maori occupy similar acoustic space, they sound the same, and therefore they can be used to classify new unknown speech units or words.

Keywords: Bilingual English and Maori speech and phonological units; acoustic segments; connectionist-based evolving clustering; speech classification

1. Introduction

The core principles of speech recognition, namely pattern recognition and classification are both generic knowledge engineering problem-solving tasks. These tasks are related to recognizing and/or classifying an unknown property or fact with a set of existing known properties or facts [1]. This study looks at the classification problem to model both New Zealand English (NZE) and New Zealand Maori (NZM) speech into known theoretical acoustic clusters. The first section presents data analysis constructed into two working

[1] *Corresponding author.* Dr Mark R. Laws, AUT Tech Park, 581-585 Great South Rd, Penrose, Auckland.
E-mail addresses; mlaws@aut.ac.nz, rkilgour@navman.com, nkasabov@aut.ac.nz.

models that can be used as the foundations for future speech recognition or classification systems. The analysis focused on the word segments extracted from speech examples from the NZE and NZM languages; this determined a baseline model. The second section discusses a novel approach to collectively label both languages with an arbitrary notation called ‘minimal acoustic segments’ (MAS). This approach will be presented as an alternative to the traditional linguistic nomenclature. The next section will briefly describe a hybrid connectionist-based approach, and the ‘evolving clustering method’ algorithm (ECM). The final section reports on the initial results of the experimental connectionist-based ‘bilingual speech clustering’ (BLSC) model.

1.1. Word Segment Analysis

We have substantial information about NZE and NZM phonology, the lexicon and the current status in terms of their bilingual relationships. Therefore we have implicit knowledge about the similarities and dissimilarities between the two languages. Thus one can make some general and informed assumptions about both. To lay down the benchmark for this BLSC problem, we first analyzed both lexicons for phoneme matching pairs to extract empirical knowledge about the data before we presented it to the connectionist architecture. Understanding the relationships between the phonemes in either language was not so important as actually knowing how the many occurrences of each appeared in their respected pairs. This will become more apparent when the results of the BLSC analysis is reviewed in Section3.2.

A training data set from NZM and NZE were assembled. The data sets contained 100 words each (see Tables 1 and 2). A testing data set was also assembled, containing additional word entries not included in the training set. All speech and word examples were extracted from a single male speaker with utterances in both languages from the MOOSE database [2].

To construct the training data sets, random samples of the NZE and NZM words were taken. Once pre-processed, these sets showed too many imbalances with the phoneme and syllable sizes between the languages. That is, there was substantially more NZM data for the same number of NZE words. Tables 1 and 2 clearly show that the NZM data set has about twice as many syllables and a third more phonemes. Syllable boundaries and the number of phonemes for each data set were determined because this was considered an important factor to note prior to the BLSC analysis [3].

The NZE 100 words contained sixty-five single syllable words and seventy double syllables, a total of 135 syllables and 311 phonemes. In NZM, there were only three single syllable words, one hundred and twenty double syllables, sixty-six with three syllables, forty-four with four and twenty with five syllables, a total of 253 syllables. There were also a total of 435 NZM phonemes. The large differences between the two languages clearly shows that the phonemic and morphemic structures are vastly different [2]. Therefore we were very interested in knowing how the under-represented individual number of NZM phoneme units (19) with its over-populated phoneme and syllable count, would compare to NZE (45 phonemes). In this respect, NZE is the opposite of NZM, with the exception of the total phoneme count.

We looked at the speech data from an acoustic perspective, and disregarded the phonological order. That is, the acoustic representation for each phoneme is separate in each language, but identical across the languages, where matches occurred. For example, ‘go’ [gou] and ‘*ahau*’ [ahou] (my) use the same NZE diphthong. With one exception, the voiced post-alveolar approximant represented as /r/ in NZE is transcribed as an alveolar trill [{] for NZM [2]. Therefore, the rolling of /r/ in NZM is very distinct compared to NZE, thus it was considered important to identify these two phonemes separately. The NZM samples also has twice as much silence before and after each word. Note the IPA transcriptions were used to identify all phonemes [4].

Phoneme frequency analysis was then performed on both lexicons to determine the number of occurring phonemes in each. In Table 3, the consonants which have high frequency counts are the shared phonemes between NZE and NZM, this includes the vowel phonemes /e/ /i/ /u/ /ç/ and /a/ which again are the five vowels in NZM. Figs. 1 and 2 give a better indication of the influence that NZM has on the overall frequency.

In summary, this data analysis was an important step before we submitted the speech data set examples to the BLSC networks. We now have data sets that represent a random sample for training and testing, but more importantly, these sets represent acoustic variations between the two languages. Because we are only interested in language variations in this preliminary study, we achieved this by using a single native speaker of both languages, thus ensuring any speaker variances were controlled.

1.2. Minimal Acoustic Segment Annotations

Linguistic annotations cover the descriptive and analytical transcriptions of a wide range of text and speech notations. Speech and language databases are one example of the linguistic annotations that have been developed to specifically describe all the elements and their relationships within this technological tool [5]. Standardization is one area of linguistic annotation that has been difficult to maintain, granted the wide fields that this format methodology must cover. As various attempts have already been made to standardize computer file formats, there have also been just as many efforts placed on the logical structures of linguistic formats [6]. In addition, a specialized view closely associated with this analysis, also uses a logical linguistic notation to describe rule node activations from a neural network classifier [7].

Therefore, the novel approach to collectively label NZE and NZM with an arbitrary annotation is based on the assumption that approximately 75% of the phonological units in both languages occupy similar linguistic and acoustic space. We have called this annotation, 'minimal acoustic segments' (MAS) because they represent sub-units of sound that can have no limit on length or quantity. A succession of MAS would represent higher orders of sound, and will appear at all levels of the classification task. The successive similarities between MAS at those higher levels could be equated to say, phonemes, syllables or even whole words. Because, initially, no assumptions were made about the BLSC analysis results would present, we could only initially say that the MAS would at the least, represent varying units of sound as output values.

To identify these units, we have decided to move away from the standard linguistic annotations and rely on this arbitrary concept during the initial labeling process. This hypothesis will be tested in Section2.2 and Section3.2 with further comments and discussions that should justify the use of a new form of annotation.

1.3. Connectionist-Based Approach

A paradigm shift from the statistical modeling approach to a symbolic rule-based approach has occurred over the past twenty years. In 1989 there was an article about connectionist systems breaking through the problem barriers with more advances than the traditional statistical modeling techniques [8]. Over this time, continued research has developed a wealth of proven scientific and engineering approaches to solve the endless problems associated with the 'artificial intelligence' (AI) and connectionist archetype. Nevertheless, both old and new approaches still have their theoretical and practical applications in this expanding field. This is also pertinent when hybrid connectionist and classical approaches are both combined to extract the best features from each method, thus adding to an even greater selection of AI connectionist problem-solving applications.

Artificial neural networks (or ANN) are dedicated parallel processing structures that can represent symbolic knowledge [9]. Many different types and models of the ANN have been applied to numerous problem-solving tasks, from classical to novel approaches [10]. Collectively, all these ANN's are known as machine-learning architectures which are generally called ‘connectionist systems’ [1]. The connectionist approach represents knowledge-based structures that can be trained with data. They can therefore learn from past experiences, they can adapt, they are able to generalize, they are robust and they utilize massive parallel connections to best apply their problem-solving capabilities [7].

1.3.1. Evolving Connectionist Systems. The Evolving Connectionist System (ECOS) toolbox is a hybrid connectionist-based system that utilizes fuzzy logic inference and evolutionary programming. ECOS was developed under the research program ‘Connectionist-based Intelligent Information Systems (CBIIS) [7].

The ECOS principles are based on connectionist structures which can be quickly modified on-line through interactions with the environment they are exposed to. ECOS will enhance the general ANN model by improving training and learning, by increasing memory and storage, and by reducing local minima problems during training [7, 11, 12, 13].

The ECOS framework is an emerging research theme which will ultimately consist of the following characteristics [14];

- Fast learning through ‘one-pass’ training;
- Adaptable to new data, features and classes;
- Open-ended structure that can accommodate for new inputs, nodes, outputs, and connections;
- Memory stored in new connections to reduce ‘catastrophic forgetting’;
- Continuously interacting with its data environment;
- Ability to explain its behavioral learning;
- Represents spatial-temporal concepts.

1.3.2. Evolving Clustering Method (ECM). An evolving clustering method (ECM) can be employed in both on-line and off-line evolving connectionist models. ECM can effectively learn complex temporal sequences in an adaptive way. The ECM and its extension, an ‘Evolving Clustering method with Constrained Minimization’ (EC-CM), both of which are used in the ECOS model for partitioning the input space [15, 16].

The evolving clustering with constrained minimizing algorithm uses the Evolving Clustering (EC) procedure in a fast, distance-based, one-pass clustering process. Optimizing of the on-line estimation for clusters and cluster centers is required, as the Constrained Minimizing (CM) optimizer processes the results from the EC to allow more suitable off-line tasks. This is in order to partition the input space with training data for creating fuzzy rules if and when required.

EC is specially designed for on-line modes employed by the ECOS framework [15]. The ECM algorithm comprises of two functions. The first function is for clustering the data set in either an on-line or off-line training mode. This function takes two parameters. The first parameter is the data to be clustered and the second (which is optional) is a list of parameters. This function returns a structure containing a field of clustered centers. This represents the network weights, and is the basis of the visualization techniques described in Section 2.1. The second ECM function is used for plotting the results of the network weights [3, 16].

ECM has been compared with other clustering methods, such as fuzzy C-means [17], and subtractive clustering method [18]. The results can outperform these well known methods [16]. The ECM algorithm is implemented in the 'Matlab' numeric computing environment [19].

2. Nonlinear Acoustic Modeling

The nonlinear acoustic modeling is the principle feature of this analysis, it uses the ECM method as the basis of the following experiments. Here we will present results based on the NZE and NZM speech data sets. All the speech examples were extracted from a single male speaker containing word utterances in both languages. The justification for using a single speaker in these initial experiments was because we were primarily interested in the language variations, as opposed to speaker variations. Three experiments are of interest, and will be compared with each other;

i) Experiment-1 involved presenting all the NZE speech data to the ECM, followed by all the NZM speech data;

ii) Experiment-2 was similar, except that the NZM speech was presented first;

iii) Experiment-3 saw both NZE and NZM data randomly shuffled together before being presented.

This analysis attempts to train three ECM models to respond differently to three controlled speech stimuli. This process was framed around the hypothesis of how learners of different languages cluster similar sounds in perceptual space. If a foreign-language sound is similar, but not the same, to a native-language sound, then the learner will categorize the sound in their native-language. When they are exposed to two languages in various orders, their ability to discriminate is very high, especially in infants [20].

In effect, we are attempting to model two types of bilinguals; the first is the adult ‘second language learner’ (NZE + NZM or NZM + NZE), and the second is the infant bilingual native-language learner (NZE- NZM mixed). Although both language models can cluster the two languages, the main learning effects are different. The first could be using the perceptual magnet effect where sounds are clustered around a winning prototype, and the second uses perceptual space to discriminate between similar sounds [21]. Ultimately, both models cluster the sounds in a similar fashion.

2.1. Speech Data Processing, Training and Visualization

The speech data is processed using a Matlab ‘GetData’ function which can take up to four parameters, only the first is mandatory. The parameters are; the ‘wave file name’; the amount of ‘overlap’ between frames; the desired ‘length’ of the sample; and the number of samples to ‘skip’. The overlap defaults to 256 samples (50%). The length defaults to the length of the entire speech file. The number of samples to skip defaults to zero. In normal usage, the last two parameters can be left as the default. The GetData function also requires the ‘readwav’ and ‘melcepst’ functions from the ‘Matlab VoiceBox’ [19].

The following function will return 512 FFT frames transformed into 26 ‘mel-scale cepstrum coefficients’ (MSCC) resulting in twelve frequency components. In addition, the log power MSCC was also calculated, thus each frame consists of twenty-seven vectors with consecutive frames overlapping by half [3]. If required, the raw wave data may also be returned.

```
function [msccdata, wavedata] = GetData(wavefilename,overlap,length,skip);
  if nargin<3,length = -1;end;
  if nargin<4,skip=0;end;
  [wavedata,FS] = readwav(wavefilename,'s',length,skip);
  if nargin<2, overlap=256;end;
  p=floor(3*log(FS));
  n=pow2(floor(log2(0.03*FS)));
  msccdata = melcepst(wavedata,FS,'Na0ye',27,p,n,overlap);
return;
```

This speech data transformation processing format has been found to be the best suited process for this type of network and classification problem [6]. Therefore the MSCC’s were presented to the ECM as a consecutive stream of unsupervised training speech examples. The first ECM function required a distance threshold (*Dthr*) of ‘0.155’ which was used for all the experiments. This threshold was determined to yield the correct amount of cluster nodes that can be represented as the arbitrary annotation of the MAS.

The transformation results for 100 NZE words produced 5,725 MSCC samples in a 27 dimensional configuration. The small distance threshold parameter keeps the number of clusters down to 61, which is a comparable total. The training data is only presented once (Epochs = 0) and the time taken for training was less than 10 minutes (see Fig. 3).

The results for 100 NZM words produced 6,832 samples (1000 more than NZE). There were 71 clusters which was similar to NZE and the training time was also slightly less (see Fig. 4).

After training the cluster plots of both languages did not seem to indicate any relative representation of actual words being presented to the network. Otherwise, the important feature here was the cluster learning rates, or evolving clusters. As the amount of MSCC samples increases, this implies that more clusters are being created to accommodate for the new unknown data examples. The learning gradient starts off steep and then begins to plateau after about 1000 samples, the learning curve then becomes more linear. This is understandable, given that new rule nodes will be created early to represent all the new presentations of the acoustic units of the language. Then as more and more units are identified as being similar to previous examples, they are aggregated and clustered to the winning neuron. Thus less and less nodes are created later on in training. There will come a point where the saturation phase of training is reached and no more new neurons are created, unless, new unknown units are presented to the network.

These initial experiments were used to test the structure of the data sets with the ECM. The results were ideal for visualization of the clusters and the evolving learning, but large areas of the cluster space was empty. This is because we were only seeing the first two dimensions of the 27-dimensional array (e.g. X1-X2). Which may not have enough relevant information to assist us in our evaluation of the ECM performance.

Therefore, an alternative statistical method was used. We decided to use the ‘principle component analysis’ (PCA) feature extraction and reduction process. This method transforms the high dimensionality of a pattern by extracting the most informative data features [22]. For example, projecting the network output weight variables (e.g. data clusters and centers) onto an alternative co-ordinate system where most of the sound unit variables will be correlated with each other, while others will not be correlated. This means the preservation of information is still maintained even when the reduction in dimensional size is performed [23]. PCA effectively finds the best features from the ‘p-dimensional’ variables and transforms them into the smaller ‘q-dimensional’ model (q = p) [24].

The first two dimensions of this co-ordinate transformation reflect the two most important extracted components. For clustered speech data, these PCA components account for approximately 20% of the variance.

Whereas the X1-X2, accounted for less than 6% of the variance. The PCA analysis on all the experiments were carried out individually. The amount of variance accounted for by the first six PCA dimensions are shown in Table 4.

Two functions allow the PCA to be used for plotting purposes [15]. The first function plots the extracted centers and data samples in PCA space. However, the dimensions of the PCA plot are now abstract, that is, they are non-linear generative models [21]. They now give us little or no clue about the new properties of the cluster centers, because they have been forced into another dimensional feature pattern. Nevertheless, we are confident that similar cluster centers will stay together, and the data points will appear closer to their appropriate cluster centers compared with the X1-X2 plots.

Two parameters are required for the 'PlotPCA' function. The first is the normalized data, the second is the structure returned from the first ECM function. Note that the returned structure contains the normalized data used in the creation of the cluster centers. However, the syntax of the PlotPCA function is intended to allow data other than that used to create the cluster centers, to be used in visualization. For example, the cluster centers found using NZE speech data may be plotted with the data extracted from the NZM speech examples. In Fig. 5 the resulting plot shows the data is well distributed.

2.2. Mono-lingual Acoustic Clustering

The second and most important function was implemented to plot a single NZE test word onto the PCA plot [2].

In Fig. 6, the start point of the word is marked with an 'x' character, and the feature points are connected with lines. The word 'nine' was plotted onto the same PCA space as in Fig. 5. Because we do not know exactly what each cluster center represents, we can only speculate about the words plot trajectory. On the middle-left of the figure, the starting point is obscured by the large area of what is assumed to be 'silence' at the beginning of the word. The first phoneme /n/ appears to be in the bottom-right hand corner and the diphthong /ai/ is at the top-right, followed by the second /n/ sound in the bottom-middle, and again returning back to the middle-left as silence in the word ending.

The final part of this speech data visualization, was to label the clusters on the PCA plots with a notation that would reflect the MAS. The only logical way in which we implemented this process was to consecutively number the clusters as they were being created. Therefore, this numeric notation gave a clear indication about the relationships between the sequences and where word plot trajectories start and end around correlating values. See Section 3.2 and Figs. 9 to 11 for an illustrated account of the MAS labeling system.

3. Modeling the Emergence of Both Languages

3.1. Bilingual Acoustic Clustering

The NZE data set (with 100 words) had created a total of 5,725 frames and NZM data created 6,832 frames. On account of the NZM data set having a greater number of syllables, this would explain the increased number of frames (see Tables 1 and 2). In Experiment-1, the NZE data alone created fifty-four clusters and an additional fourteen were created once the NZM data was presented. Experiment-2, the NZM data created forty-nine clusters and in addition, the NZE data produced fifteen more clusters. For Experiment-3, both languages produced more clusters than either of the languages did when presented separately (see Table 5).

The lower number of first language clusters for NZM then NZE is understandable, given the phoneme inventory for NZM is much smaller. Although the NZM syllable count was much higher, this has managed to balance out the cluster numbers slightly.

Fig. 7 shows the overall evolving cluster learning rates for each ECM experiment. These three gradients were tracked over the period of their total MSCC frame samples and the number of clusters created for each. Experiment-1 (NZE + NZM) shows a predictable learning curve (e.g. close to linear) for the first language, as the majority of the clusters have been created (e.g 54). Then as the NZM data was introduced to the network (e.g. at E-M), far less clusters were created for about the same amount of data samples presented. The second language learning curve overall was still linear, but not as steep. This is the ECM aggregation of learning, where ‘sounds-like’ units from the second language are mapped onto the same winning clusters (neurons) of the first language.

Experiment-2 (NZM + NZE) initially shows a similar learning curve to the first, but after about 3,000 NZM samples the aggregated learning starts much earlier. This seems to be the result of all the possible acoustic units in NZM having already been presented to the network. From this point on, there are less clusters created by comparison to the first experiment. Thus, the angle is not as steep, but the curve remains relatively linear right up until the second language was introduced (e.g. at M-E). Here we can see that once the NZE samples are presented, the curve seems to run parallel with the first experiment, with a convergence only happening near the end. One would have expected an early convergence of the two curves, with an initial creation of more clusters to allow for the new ‘unknown acoustic units’ from NZE. Nevertheless, we can assume the longer periods of aggregation are probably due to the initial slow response of the ECM to self-organize the clusters on-line. Furthermore, the time taken for the nodes in the multidimensional space to be re-organized, may have also had an effect on the number of clusters being created around these nodes.

Experiment-3 has provided the most interesting results from Fig. 7, as the random mixed set of words from both languages has revealed some contradictory results. From the onset, the learning did not follow the expected steep curve we were looking for. Granted, that having a total of seventy clusters would indicate a higher rate of learning compared to the first two experiments. As this was not the case, we speculated on the results in the following way. The random sort function may not have distributed the samples in the linear manner we would have expected across the entire sample period. The first case scenario is that one language is over represented at the beginning, but there is enough of the other language present to alter the learning pattern compared to the first two experiments. There is the possibility that many words with similar phonemes were shuffled to the beginning of the data set. There is also the possibility that many words with contrasting or opposing phonemes could have caused this poor learning initially. Furthermore, words with more silence than others may have influenced this outcome (e.g. it was noted that the NZM speech examples have much more silence than the NZE examples). This is especially evident after the twentieth neuron was created, when there were no new neurons created for well over 1,000 samples. Probably the most likely case is the larger array of complexity that both languages would create when the two divergent samples were mixed (see Fig. 1 and 2). This maybe especially true when we consider that NZE has a higher number of consonants over vowels, but with the overall greater distribution. Whereas, NZM have the same amount of consonants and vowels, at a higher frequency, but with less distribution overall.

A final reason for Experiment-3 yielding contradictory results, could be a combination of all the above. Therefore, we analyzed each of the nodes at the time they were created, and counted the number of samples that were associated with each node from the two languages. Fig. 8 clearly shows the proportions of language samples appearing at each node over the entire training period.

If we compare Fig. 7 with 8, the first obvious point is that NZM samples have swamped the initial training data right up to about the twenty-sixth node. Note that there are very few samples clustered at the twentieth node, given that the 1,000 odd samples over this period were all being aggregated. Then after the twenty-sixth node, NZE starts to become more predominant, which in turn influences the increase in learning rate on a much steeper gradient, which starts to become linear right to the end. In Fig. 7, the mixed data indicates that at this point the learning surpasses the other two experiments, and would seem to continue to evolve further. The large amount of samples appearing at nodes ‘1’, ‘28’ and ‘64’ can be assumed to be silence.

3.2. Bilingual Test Results

The next series of results shows a test word trajectory for each of the three experiments. The output bilingual plots were created using the PlotWord PCA function [2]. Here, the following series of plots are compared with each other, as the same NZE word was projected for each experiment.

In Fig. 9, the word ‘zoo’ is projected onto the NZE + NZM PCA plot. The start is marked with an ‘x’ at the upper-left, with the words trajectory points marked as connected lines. Using the MAS notation, we can say that labels ‘1 to 54’ represent shared NZE and NZM space, with labels ’55 to 68’ being the difference when NZM was presented. Note, NZM may also have had a slight effect on where the NZE labels were finally plotted. The upper-left of the plot is the starting point for silence. The phoneme /z/ appears to be in the upper-middle part and the /u/ is at the upper-right, and then the trajectory returns back through the middle to silence again.

In Fig. 10, the same word is projected onto the NZM + NZE PCA plot. The MAS ‘1 to 54’ are the shared labels for NZM and NZE with labels ’55 to 64’ being the difference when NZE was presented. The mid-left starts with silence and phoneme /z/ is in the middle with /u/ at the lower-right, then the trajectory returns back through the middle to silence again. Although the MAS labels are plotted in different PCA space compared to Fig. 9, the word trajectory here appears to be a mirror image of the word in Fig. 9. The distance is maintained but the plots are different.

In Fig. 11, the mixed language PCA plot shows a similar resemblance to Fig. 9 for the MAS labels. Again, the mid-left starts with silence and /z/ is in the middle with /u/ at the upper-right, then as usual the trajectory returns back through the middle to silence again. The word trajectory also appears similar to Fig. 9 with the distance and plots being maintained.

To summarize this section, the test results for one word projected onto three different bilingual PCA plots clearly shows that the words acoustic frames (including silence) hold their relative position of distance and clusters. Furthermore, the MAS labels can represent the sub-unit acoustic space of the speech data, which when grouped together can form phonemes of similar ‘distinctive features’ (DF), and when highly distributed, they can represent phonemes with opposing DF [2, 4].

Future experiments should include increasing the number of word examples. This should test the general rule that once all the clusters have been created for a particular language, any number of extra speech examples will be aggregated, thus the initial nodes will remain stable. Furthermore, on account that NZM has heavily influenced most of the phonemes that it has clustered with NZE, a more selective process in the number of phonemes and syllable boundary words in NZM should be undertaken. This may also account for the initial slow learning rate of Experiment-3 where NZM samples saturated the ECM.

4. Conclusion

Reporting on the connectionist-based bilingual speech clustering system was not about the acoustic classification of each language, but more about modeling the emergence of the acoustic spaces within the bilingual speech framework. As a conclusion, the evolving cluster analysis enables one language to be easily added to an existing system, along with its accents and modifications. The clustering of nodes that can represent all the acoustic units in n-dimensional perceptual space are the ideal basis for constructing future speech classification systems.

Clearly, the results point toward a future connectionist-based architecture as an effective means of classifying speech. But this comes with a cautionary note that these experiments are still in their infancy, and would require much further research and analysis (e.g. more speech examples, more speakers and other languages) to fully justify the claim that these AI architectures are effective at classifying and visualizing speech.

This paper has attempted to tackle the most difficult task of speech classification using a bilingual speech data approach. Furthermore, we had also taken a risk by using a newly developed connectionist-based method that has just recently been exposed to the public domain, and is now receiving peer reviews. Therefore, this paper in general could be viewed as using a novel approach with a novel method to solve a long-standing problem. Nevertheless, the goal to integrate a minority language, such as Maori, from the language and linguistic paradigms with an emerging information technology has now been undertaken. Although, further comparative research and development is definitely required.

Acknowledgments

The authors would like to acknowledge the Connectionist-based Intelligent Information Systems research programme which was funded by the New Zealand Foundation for Research, Science and Technology (FRST – grant UOO808).

References

[1] Kasabov, N. (1996). Foundations for Neural Networks, Fuzzy Systems and Knowledge Engineering. Cambridge, MA, MIT Press.

[2] Laws, M R. (2001) "Maori Language Integration in the Age of Information Technology: A Computational Approach." Ph.D Thesis, University of Otago. <http://kel.otago.ac.nz/maaka/phd.html>

[3] Kilgour, R. (2001). Evolving Clustering for Bilingual Speech Analysis. Dunedin, University of Otago: 18.

[4] Ladefoged, P. (1993). A Course in Phonetics.

[5] Bird, S., Liberman, M. (2001). "A formal framework for linguistic annotation." Speech Communication 33(1-2): 23-60. <http://www.elsevier.nl/locate/specom>

[6] Taylor, l., Black, A. and Caley, R. (2001). "Heterogeneous relation graphs as a formalism for representing linguistic information." Speech Communication 33(1-2): 153-174.

[7] Kasabov, N. (1999). Evolving Connectionist and Fuzzy-Connectionist Systems: Theory and Applications for Adaptive, Online Intelligent Systems. Neuro-Fuzzy Techniques for Intelligent Information Processing. N. Kasabov and R. Kozma. Heidelberg, Physica Verlag: 111-144.

[8] Waibel, A. and Hampshire, J. (1989). Building Blocks for Speech: Modular neural networks are a new approach to high-performance speech recognition. BYTE. August 1989: 235-242.

[9] Zurada, J M. (1992). Introduction to Artificial Neural Systems. New York, West Publishing Company.

[10] Amari, S. and Kasabov, N. (1998). Brain-like Computing and Intelligent Information Systems. Singapore, Springer Verlag.

[11] Watts, M. (1999). Evolving Connectionist Systems for Biomedical Applications. ICONIP/ANZIIS/ANNES'99 International Workshop "future Directions for Intelligent Systems and Information Sciences"., Dunedin, New Zealand, University of Otago.

[12] Kasabov, N. (1998). "The ECOS Framework and the ECO Learning Method for Evolving Connectionist Systems." Journal of Advanced Computational Intelligence 2(6): 1-8.

[13] Kasabov, N. (1998). "Evolving Fuzzy Neural Networks: Theory and Applications for On-line Adaptive Prediction, Decision Making and Control." Australian Journal of Intelligent Information Processing Systems 5: 154-160.

[14] Deng, D., Koprinski, I. and Kasabov, N. (1999). RICBIS: New Zealand Repository for Intelligent Connectionist-Based Information Systems. ICONIP/ANZIIS/ANNES'99 International Workshop "future Directions for Intelligent Systems and Information Sciences"., Dunedin, New Zealand, University of Otago.

[15] Song, Q. (2000). Evolving Clustering Method: ECCM. Dunedin, University of Otago: 4.

[16] Kasabov, N. and Song, Q. (2001) DENFIS: Dynamic evolving neuro-fuzzy inference system. IEEE Transactions on fuzzy systems, in print.

[17] Bezdek, J.C. (1981). "Pattern Recognition with Fuzzy Objective Function Algorithms." Plenum Press, New York.

[18] Chiu, S. (1994). “Fuzzy Model Identification Based on Cluster Estimation”, Journal of Intelligent & Fuzzy System, Vol. 2, No. 3, Step.

[19] Demuth, H. and Beale, M. (1996). Neural Network ToolBox For Use with MatLab. Natick, Massachusetts, The Math Works, Inc.

[20] Taylor, J., Kasabov, K. and Kilgour, R. (1999). Modelling the emergence of speech sound categories in evolving connectionist systems. 5th Joint Conference on Information Sciences, Atalantic City, NJ. <http://kel.otago.ac.nz/maaka/index.html>

[21] Kuhl, Patricia. (1991). “Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not.” Perception & Psychophysics 50(2): 93-107.

[22] Cios, K., Pedrycz and Swiniarski (1998). Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers.

[23] Westphal, C. and Blaxton, T. (1998). Data Mining Solutions: Methods and Tools and Techniques for Solving Real-World Problems, John Wiley and Sons Inc.,.

[24] Pal, N R. (1998). Connectionist Approaches for Feature Analysis. Brain-like Computing and Intelligent Information Systems. S. Amari and N. Kasabov. Singapore, Springer Verlag: 147-167.

Appendix.

Table 1
The 100 selected NZE words, showing the phonetic transcriptions and the syllable count for each word.
Note the number of syllables in total is 135

100	IPA	Number	100	IPA	Number
NZE Words:	Transcriptions:	Syllables:	NZE Words:	Transcriptions:	Syllables:
ago	´ g ou	2	nod	n Å d	1
ahead	´ h e d	2	one	w v n	1
air	E´	1	ooze	u z	1
any	e n i	2	other	v D ´	2
are	a r	1	over	ou v ´	2
auto	ç t ou	2	palm	p a m	1
away	´ w ei	2	paper	p ei p ´ r	2
baby	b ei b i	2	pat	p Q t	1
bat	b Q t	1	pea	p i	1
bird	b Œ d	1	peace	p i s	1
boo	b u	1	pure	p j u´	2
book	b U k	1	push	p r U S	1
boot	b u t	1	rather	r a D ´ r	2
buzz	b v z	1	recent	r i s I n t	2
card	k a d	1	reef	r I f	1
carrot	k Q r Å t	2	riches	r I tS I s	2
choke	tS ou k	1	river	r I v ´ r	2
coffee	k Å f i	2	rod	r Å d	1
dart	d a t	1	rouge	r ou Z	1
day	d ei	1	rude	r ou d	1
dead	d e d	1	school	s k u l	1
die	d ai	1	seven	s e v I n	2
dog	d Å g	1	shoe	S u	1
dove	d v v	1	shop	S Å p	1
each	i tS	1	sing	s I N	1
ear	i´	1	singer	s I N Œ	2
eight	ei t	1	six	s I k s	1
ether	i D ´	2	sue	s ou	1
fashion	f Q S n	2	summer	s v m ´ r	2
fat	f Q t	1	tan	t Q n	1
five	f ai v	1	tart	t a t	1
four	f ç r	1	teeth	t i T	1
fur	f Œ	1	teethe	t i D	1
go	g ou	1	that	D Q t	1
guard	g a d	1	thaw	T ç	1
gut	g v t	1	there	D E´	2
hat	h Q t	1	three	T r i	2
hear	h i´	2	tour	t u´	2
how	h aw	1	tragic	t r Q dZ k	2
jacket	dZ Q k I t	2	tub	t v b	1
joke	dZ ou k	1	two	t u	1
joy	dZ çi	1	utter	v t ´ r	2
judge	dZ v dZ	1	vat	v Q t	1
lad	l Q d	1	visit	v I z I t	2
ladder	l Q d ´	2	wad	w Å d	1
leisure	l e Z	2	yard	j a d	1
letter	l e t Œ	2	yellow	j e l ou	2
loyal	l çi l	2	zero	z i r ou	2
mad	m Q d	1	zoo	z u	1
nine	n ai n	1	zoo	z u	1

Table 2
The 100 selected Mäori words, showing the phonetic transcriptions and the syllable count for each word.
Note the number of syllables in total is 253, nearly twice as many as NZE

100	IPA	Number	100	IPA	Number
Mäori Words:	Transcriptions:	Syllables:	Mäori Words:	Transcriptions:	Syllables:
ahakoa	a h a k ç a	5	nama	n a m a	2
ahau	a h ou	2	nei	n ei	1
ahiahi	a h i a h i	4	ngahere	N a h e { e	3
ake	a k e	2	ngaro	N a { ç	2
ako	a k ç	2	ngäkau	N a k ou	2
aku	a k u	2	ngäwari	N a w a { i	3
anake	a n a k e	3	ngeru	N e { u	2
anei	a n ei	2	noa	n ç a	2
anö	a n ç	2	nöu	n ç u	1
aroha	a { ç h a	3	nui	n u i	2
atu	a t u	2	oma	ç m a	2
atua	a t u a	3	one	ç n e	2
aua	ou a	2	oneone	ç n e ç n e	4
auë	ou e	2	ono	ç n ç	2
ähei	a h ei	2	ora	ç { a	2
ähua	a h u a	3	oti	ç t i	2
ähuatanga	a h u a t a N a	5	otirä	ç t i { a	3
äkuanei	a k u a n ei	4	öku	ç k u	2
äpöpö	a p ç p ç	3	öna	ç n a	2
ätähua	a t a h u a	4	pakaru	p a k a { u	3
äwhina	a f i n a	3	pekepeke	p e k e p e k e	4
engari	e N a { i	3	pikitia	p i k i t i a	4
ëhara	e h a { a	3	poti	p ç t i	2
ënä	e n a	2	putiputi	p u t i p u t i	4
ënei	e n ei	2	rangatira	{ a N a t i { a	4
ërä	e { a	2	reri	{ e { i	2
ëtahi	e t a h i	3	ringaringa	{ i N a { i N a	4
hanga	h a N a	2	rongonui	{ ç N ç n u i	4
heke	h e k e	2	rüma	{ u m a	2
hine	h i n e	2	tahuri	t a h u { i	3
hoki	h ç k i	2	tënä	t e n a	2
huri	h u { i	2	tikanga	t i k a N a	3
iho	i h ç	2	tokorua	t ç k ç { u a	4
ihu	i h u	2	tuna	t u n a	2
ika	i k a	2	unu	u n u	2
inäianei	i n a i a n ei	5	upoko	u p ç k ç	3
ingoa	i N ç a	3	uri	u { i	2
inu	i n u	2	uta	u t a	2
iwa	i w a	2	utu	u t u	2
kaha	k a h a	2	waha	w a h a	2
kaiako	k ai a k ç	3	waiata	w ai a t a	3
kete	k e t e	2	wero	w e { ç	2
kino	k i n ç	2	whä	f a	1
koti	k ç t i	2	whakahaere	f a k a h a e { e	5
kura	k u { a	2	whakapai	f a k a p ai	3
mahi	m a h i	2	whanaunga	f a n ou N a	3
mea	m e a	2	whero	f e { ç	2
miraka	m i { a k a	3	whiri	f i { i	2
motu	m ç t u	2	wiki	w i k i	2
muri	m u { i	2	würu	w u { u	2

Table 3
The 46 phonemes for NZE and Mäori, showing the frequency count for each phoneme

Corpu sCode:	Phoneme:	NZE Count:	Mäori Count:	Total Count:	Corpu sCode:	Phoneme:	NZE Count:	Mäori Count:	Total Count:
1	p	9	11	20	23	w	3	7	10
2	b	9	-	9	24	j	3	-	3
3	t	26	23	49	25	I	12	-	12
4	d	19	-	19	26	e	7	31	38
5	k	10	29	39	27	Q	14	-	14
6	g	5	-	5	28	v	9	-	9
7	f	7	7	14	29	A	7	-	7
8	v	7	-	7	30	V	2	-	2
9	T	3	-	3	31	i	12	43	55
10	D	6	-	6	32	a	8	95	103
11	s	11	-	11	33	ç	3	36	39
12	z	6	-	6	34	Œ	4	-	4
13	S	4	-	4	35	u	8	34	42
14	Z	2	-	2	36	ei	5	6	11
15	h	4	24	28	37	ai	3	3	6
16	tS	3	-	3	38	çi	2	-	2
17	dZ	6	-	6	39	ou	11	5	16
18	m	3	8	11	40	au	1	-	1
19	n	9	27	36	41	i´	2	-	2
20	N	2	15	17	42	u´	2	-	2
21	l	8	-	8	43	Œ´	2	-	2
22	r	20	-	20	45	´	12	-	12
50	{	-	31	31					

Table 4
Variance accounted for by various numbers of PCA dimensions

# PCA	Variance for each language		
Dimensions:	NZE:	Mäori:	Both:
1	0.1216	0.1299	0.1084
2	0.1821	0.1956	0.1643
3	0.2229	0.2315	0.2024
4	0.2587	0.2639	0.2315
5	0.2852	0.2940	0.2552
6	0.3081	0.3172	0.2755

Table 5
The number of clusters created for each experiment

Experiments:	Number of first language clusters:	Total number of clusters:	Difference:
NZE then Mäori:	54	68	14
Mäori then NZE:	49	64	15
Both Languages Mixed:	-	70	-

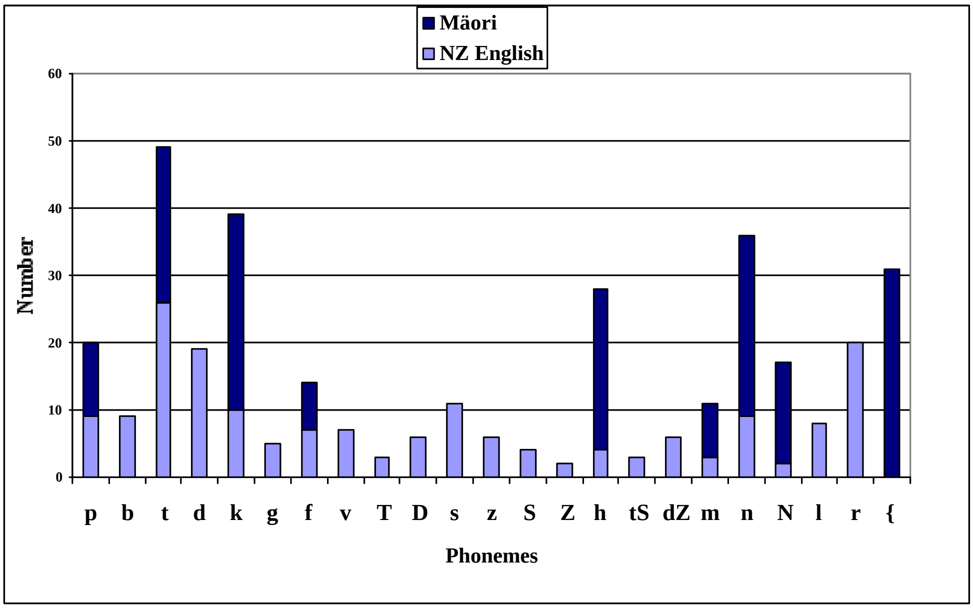


Fig. 1. NZE and Mäori consonant phoneme counts from100 words each.

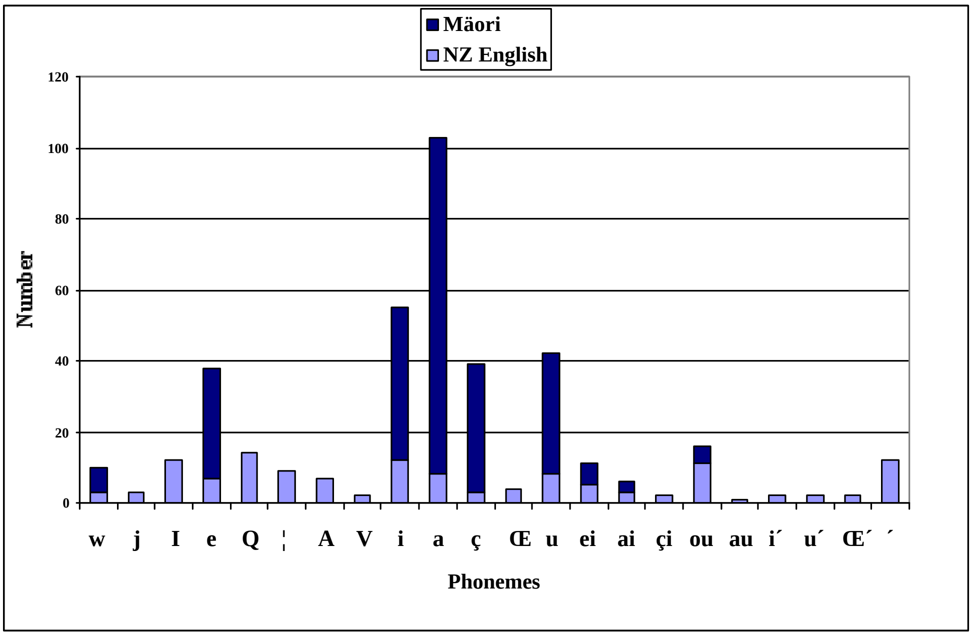
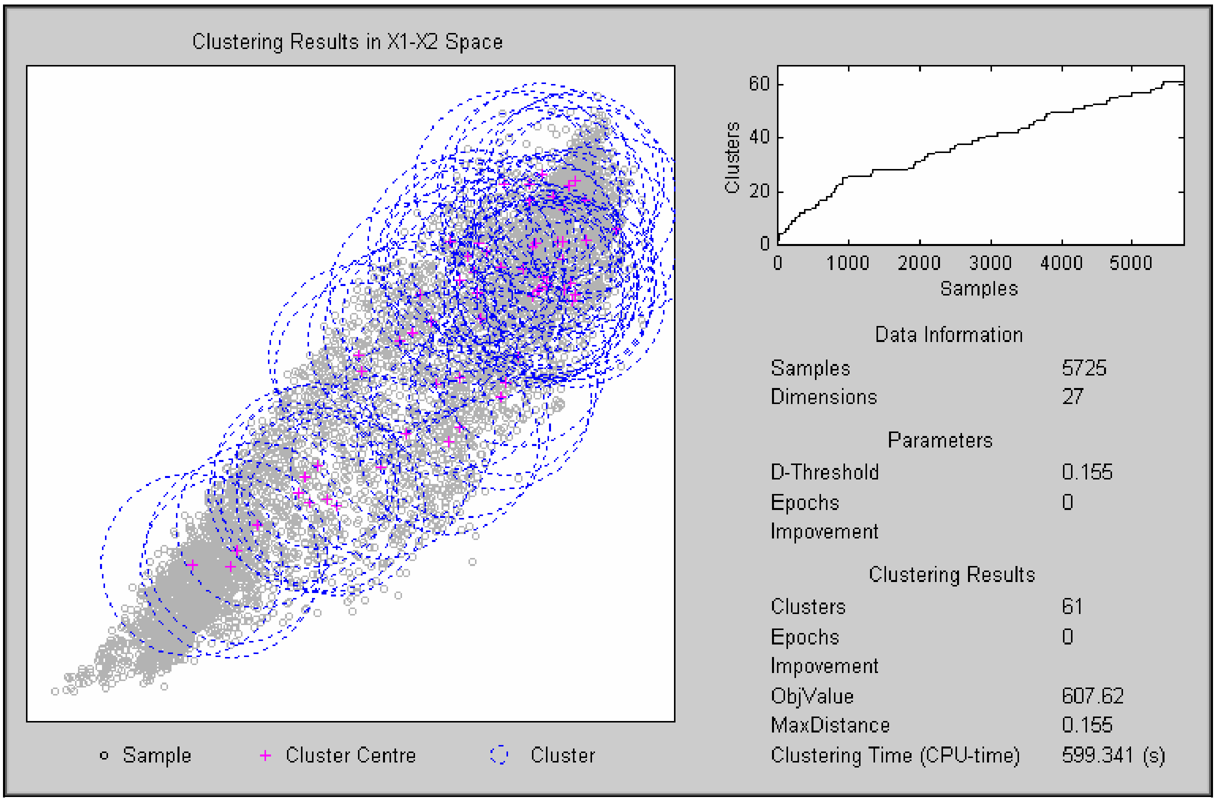


Fig. 2. NZE and Mäori vowel phoneme counts from 100 words each. Note that although /w/ and /j/ function phonotactically as consonants, they are acoustically equivalent to vowels and are therefore included in the vowel counts.



Data Information

Samples	5725
Dimensions	27

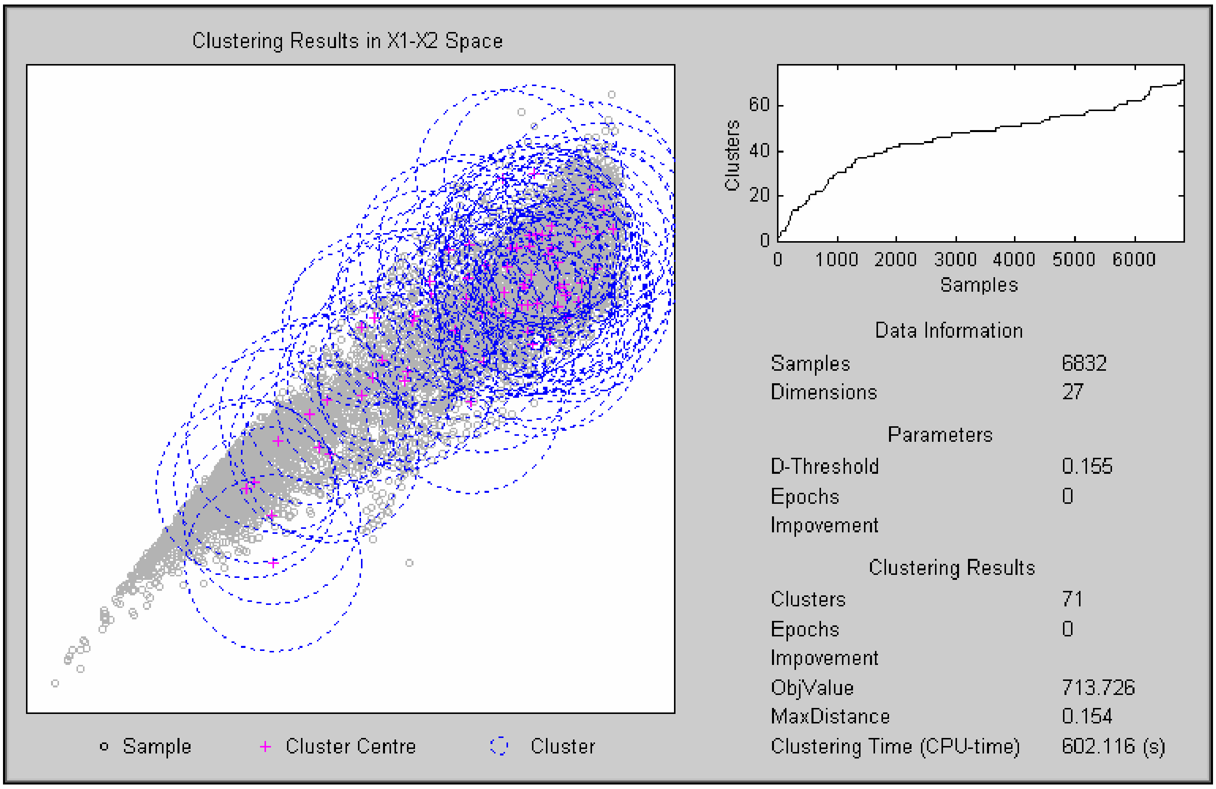
Parameters

D-Threshold	0.155
Epochs	0
Impovement	

Clustering Results

Clusters	61
Epochs	0
Impovement	
ObjValue	607.62
MaxDistance	0.155
Clustering Time (CPU-time)	599.341 (s)

Fig. 3. Plotting the custering results for 100 NZE words. Showing 5,725 MSCC samples in a 27 dimensional configuration, of which only the first two are plotted (X1-X2). The small distance threshold parameter keeps the number of clusters to a comparable total. Also note that the training data is only presented once (Epochs = 0) and the time taken for training is less than 10 minutes.



Data Information

Samples 6832
Dimensions 27

Parameters

D-Threshold 0.155
Epochs 0
Impovement

Clustering Results

Clusters 71
Epochs 0
Impovement
ObjValue 713.726
MaxDistance 0.154
Clustering Time (CPU-time) 602.116 (s)

Fig. 4. Plotting the custering results for 100 Mäori words. Note even though there are more samples for Mäori (1000+), the clusters are similar to the NZE and the training time is slighly less.

Fig. 5. PCA plot results for 100 NZE words. The grey mass are all the data samples and the black circles represent the cluster centres.

Fig. 6. The NZE word ‘nine’ plotted in PCA space. Note if we discount silence, there are three main trajectory points indicating the start of the word, the middle and the ending.

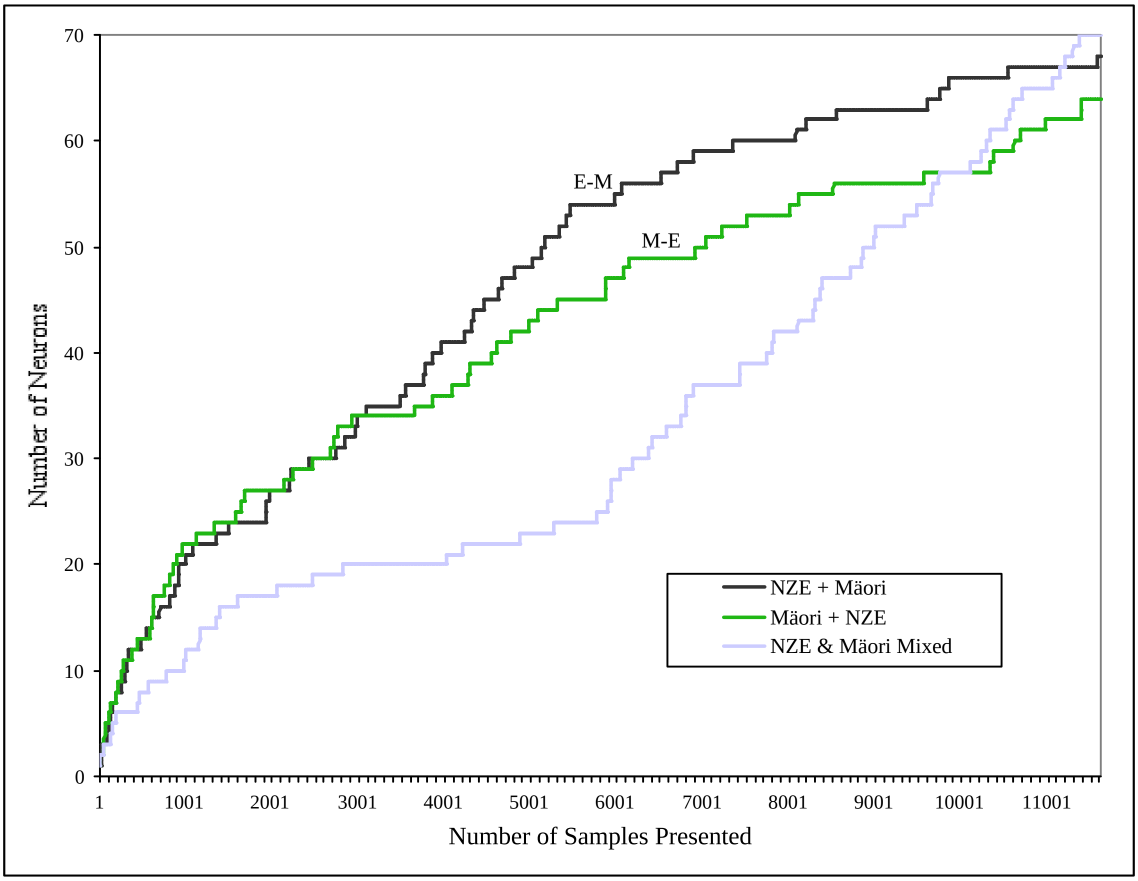


Fig. 7. Plots the evolving clusters over the entire sample periods for all three experiments. Note the ‘E-M’ and ‘M-E’ positions which indicate where one language ends and the other starts.

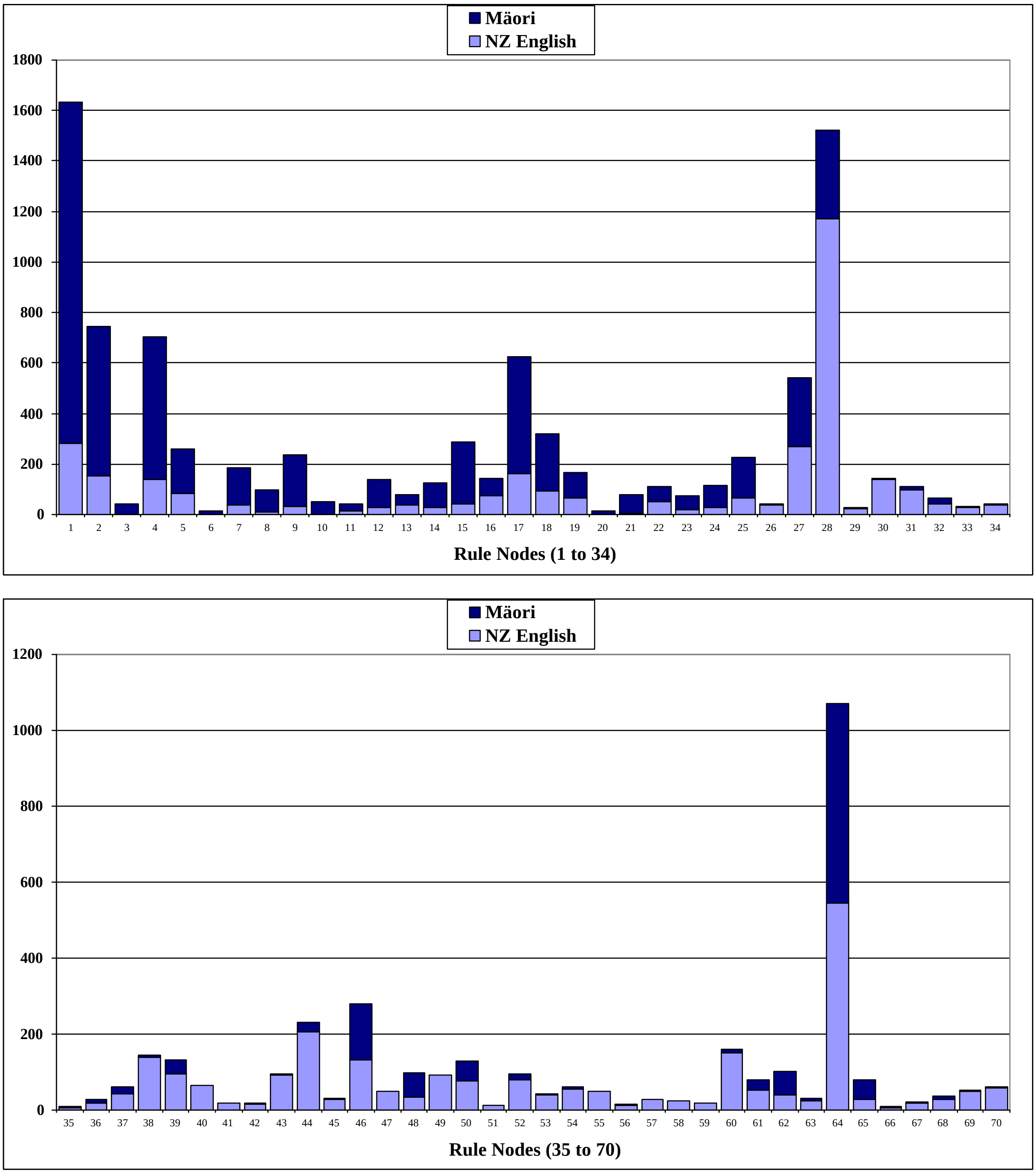


Fig. 8. Mixed proportion of language over the number of samples at each node.

Fig. 9. NZE + Mäori MAS labelled plots on the PCA space, with the word ‘zoo’ projected.

Fig. 10. Mäori + NZE MAS labelled plots on the PCA space, with the word ‘zoo’ projected.

Fig. 11. Mixed language MAS labelled plots, with the word ‘zoo’ projected.