Proceedings of the Institute of Acoustics

TWO DIMENSIONAL REPRESENTATION OF PHONEMES OF THE ENGLISH LANGUAGE

E.M. Ellis and A.J. Robinson

Cambridge University Engineering Department Trumpington Street, Cambridge, CB2 1PZ.

1. INTRODUCTION

The relationship between any two phonemes of a language is a complex one governed by a number of different parameters. Phoneticians have identified many parameters such as the place and manner of articulation of general speech, and are able to categorise the phonemes using these (Flanagan [1]; O'Connor [2]). These parameters are commonly referred to as features. The sorts of parameters described above are useful for placing the phonemes into the different classes such as vowels, fricatives, stops etc., but the multi-dimensional parameter representation makes it difficult to analyse the phoneme set as a whole, not clearly showing how the different phoneme classes may be related.

To represent the phonemes collectively it is clear that a different method of analysis and classification is needed. It would be desirable to reduce the multi-dimensional phoneme representation to a two or three-dimensional pattern that can be analysed and quantified visually. The ultimate aim of this study is to illustrate the phonemes on a two-dimensional format in a logical and natural way, and is to be achieved with minimal loss of information. A way to achieve this is to represent the phonemes according to their similarity, maintaining as much of the original parameter information as possible. The resulting representation is not intended to show features or any other such relational parameters, but will illustrate, as a Euclidean distance measure, how similar the phonemes of the English language are. That is to say, similar sounding phonemes will appear close together on a two-dimensional format, whereas dissimilar phonemes will be spaced far apart. This method of representation is intended to encompass as much parameter information as possible and display it collectively on a low dimensional format.

From an acoustic level the similarity between phones can be determined by considering the first and second formant frequencies. Earlier work by others (Rabiner and Schafer [4]) have produced good results, but have generally been confined to the vowels and semivowels. Other methods include extracting key features from the speech signal such as total energy content, number of zero crossings etc., often used as the input parameters for neural based systems (Kohonen [5]; Robinson [6]). One presentation investigated by Shepard [7], uses a second order gradient decent method on a set of manually obtained confusion data (correct and incorrect responses from a subject when presented with a random set of spoken phones). These are examples of the numerous methods available for performing a proximity analysis of this kind. Although many do not operate on the full phoneme set, the general results from these methods are found to be much the same.

The proposed analysis can be useful as a handle when considering the performance of a speech recognition system. The proximity information is a comprehensive illustration of how the phonemes are confused, and can be incorporated to reduce the redundancy in recognition techniques. For

example, to highlight the classes or class-boundaries that introduce the most recognition errors, hopefully to improve overall performance.

The two-dimensional format also lends itself quite naturally to the transmission of speech information via tactile means for the hearing impaired.

2. METHOD OF ANALYSIS

The system for analysing the phonemes can be broken down into a small number of stages. The general form of this speech analysing system is illustrated in figure 1.

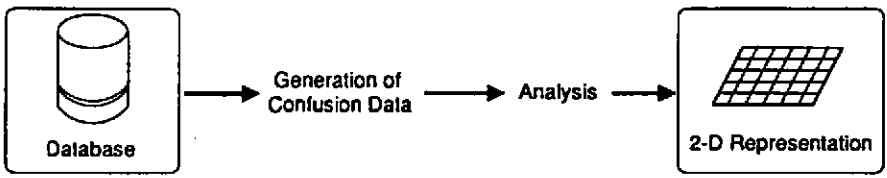

Figure 1: The Complete System

The database used to feed the system can take on many different forms but is essentially a collection of phones stored in a format that can be reconstructed by some means to give the original acoustic signals representing the phonemes. The simplest form of a database would be a collection of randomly chosen phones recorded acoustically; for this case the confusion data is constructed by having a subject listen to the recorded data and make a decision as to what phoneme was perceived. The errors in making this decision will be the source of generating confusion data.

Generally speaking, the database is a collection of indistinct elements from which confusion information can be generated. The confusion data will almost certainly be represented as a matrix (see David and Denes [7]; Green [8]) where the errors in recognition or classification are shown as the off-diagonal elements. To support this, the confusion matrix representation is a common form of input to the proximity analysis methods available to generate a low dimensional representation of the phonemes.

Various established methods exist for discriminately searching a multi-dimensional space occupied by confusable elements (Green [8]; Duda and Hart [9]), and many of these methods will attempt to reduce the number of dimensions whilst maintaining a high percentage of the original confusion information. The analysis stage has the task of reducing the dimensionality of the phoneme representation down to just two, to allow the planar format that is desired.

3. PRACTICAL IMPLEMENTATION

3.1 Generating Confusion Data

The phoneme recognition and subsequent generation of confusion data has been implemented with a recurrent error propagation network phoneme recogniser (Robinson [6]) and is the practical means by which the general results in this paper have been produced. The recurrent net recogniser has been trained and tested on the DARPA TIMIT Acoustic Continuous Speech Database (hereafter referred to as the TIMIT database) which is a well recognised standard (Lamel *et al.* [10]). Using a well recognised database makes comparisons with other speech analysing systems less difficult. The complete TIMIT CD-ROM (1990) is a large vocabulary, multiple speaker database, consisting of 630 speakers uttering 6300 sentences. The recognition accuracies achieved from the recurrent error propagation network based on the TIMIT database are around the 75% mark (Robinson [6]), and it is the remaining errors in recognition that allow confusion data to be generated. Table 1 shows the phonemes that are used with the TIMIT database along with a word in which the phoneme appears. All results will be based on this phoneme set.

Vowels		Semivowels		Nasals	
eh	bet	l	lat	m	mom
ih	bit	el	bottle	em	bottom
ao	bought	r	ray	n	noon
ae	bat	w	way	en	button
aa	bott	y	yacht	ng	sing
ah	but	hh	hay	eng	washington
uw	boot	hv	ahead	nx	winner
uh	book				
er	bird	Fricatives		Stops	
ux	toot	s	sea	p	pea
ay	bite	sh	she	b	bee
oy	boy	z	zone	t	tea
ey	bait	zh	azure	d	day
iy	beet	th	thin	k	key
aw	bout	dh	then	g	gay
ow	boat	f	fin	dx	muddy
ax	about	v	van	q	bat
axr	butter				
ix	debit	Affricates		Closures	
ax-h	suspect	ch	choke	pcl	pea
		jh	joke	bcl	bee
Others				tcl	tea
h#				dcl	day
pau				kcl	key
epi				gcl	gay

Table 1: Phonemes of the TIMIT Database

3.2 Analysis by Principal Components

Principal Component Analysis is one method for performing a proximity analysis and is the method to be used to determine the feasibility of representing the phonemes of table 1 on a low dimensional space (i.e. two dimensions). The method (described more fully by Green [8]; Hertz *et al.* [11]) searches for the maximum variance of a number of confusable elements in a multi-dimensional space, and generates a number of vectors which point in the directions of maximum variance. The method is achieved through a process of generating eigenvalues and eigenvectors from a confusion matrix of the input data.

The output of the phoneme recogniser is a vector which shows the most likely phoneme that was detected in that time frame and a measure of how it is confused with the other phonemes. For an output vector $\underline{y}(t)$ a mean compensated confusion matrix is generated for a large number of input frames using,

$$C=E\{[\underline{y}(t)-\underline{\mu}(t)]\,[\underline{y}(t)-\underline{\mu}(t)]^T\} \qquad \text{where} \quad \underline{\mu}(t)=E\{\underline{y}(t)\}$$

and C is known as the covariance matrix.

The eigenvectors of the covariance matrix will point in the directions of maximum variance. The eigenvector belonging to the largest eigenvalue will lie in the plane of maximum variance. The next largest eigenvalue will be associated with an eigenvector orthogonal to the first and in a direction of maximum variance. Further orthogonal vectors will be created in this manner for as many dimensions in the original space. The idea is to capture a large percentage of the overall variance of the input data in the first few principal dimensions (two in the case of a planar representation), to justify disregarding the other components.

The resulting analysis provides distance measures of the different phonemes according to how similar they are for the principal components that are of concern. Considering just the two most dominant components, a set of cartesian coordinates can be formulated and plotted. Equally, the phonemes could be displayed in three dimensions, as has been done by others (David and Denes [7]), but visually the information is not as clear and easily presentable when considering the complete phoneme set.

4. RESULTS

The analysis described in the previous section has been performed using TIMIT as the data source, and some of the key findings are discussed here.

4.1 All Phoneme Classes

Figure 2 is a plot of the symbols of table 1 illustrating the similarity between the phonemes according to the analysis. This plot has been generated from the the two principal components only. The plot shows a reasonable amount of grouping of the phonemes according to their class, which is what was hoped. But a closer inspection reveals an area of confusion suggesting that a non-linear transformation of some description may be required to get a clearer picture.

TWO DIMENSIONAL REPRESENTATION OF PHONEMES

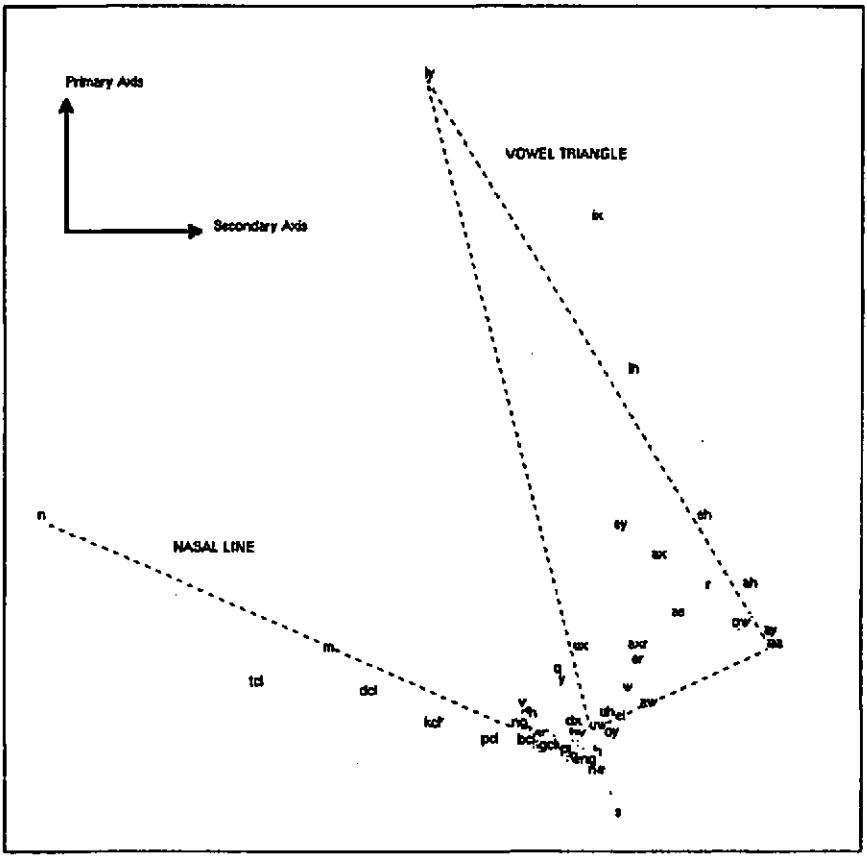

Figure 2: The Phonemes on a 2-D Space

Noticeably the vowels are not included in this area of confusion. It will further be shown, by zooming in on figure 2, that the area of confusion is due mainly to the dominance of the vowels. One feature that does manifest itself quite prominently is the well known vowel triangle, which is very much consistent with earlier work by others using different methods (Rabiner and Schafer [4]; David and Denes [7]).

Interestingly the nasals lie on a line in the two-dimensional space. This is also true for the closures (shown more clearly in the scaled version, figure 3). It is found that greater variance for these classes do exist in orthogonal planes, but the difference in variance is not significant enough to affect the general result of figure 2.

4.2 The Less Dominant Phoneme Classes

From an acoustic viewpoint the vowel, semivowel and nasal classes contain phonemes that have a high content of voicing (i.e. the vocal cords are excited for the production of these phonemes). As a result the formant frequencies feature heavily in the spectra of these phonemes (Flanagan [1]; Rabiner and Schafer [4]). Since the formant frequencies are effectively the resonant frequencies of the vocal tract they will by definition dominate the spectrum. The fricatives, affricates, stops and closures on the other hand produce weak spectra that mainly consist of high frequency components. This would be consistent with the fact that the plot of figure 2 is dominated by the phonemes which have a high content of voicing. The vowels, semivowels and nasals have been deleted from the plot of figure 2, effectively enlarging the area of confusion containing the fricatives, affricates, stops and closures. The resulting picture is shown in figure 3. Here the confusion information between the less dominant phonemes can be viewed more readily. The three 'silence' phonemes have also been removed for convenience.

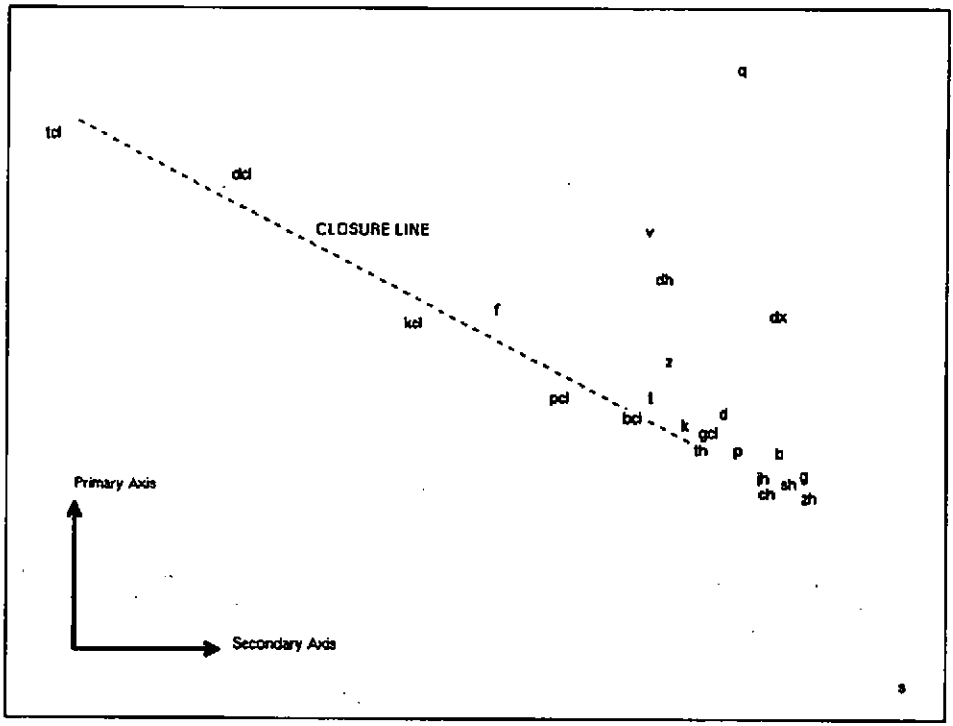

Figure 3: The Less Dominant Phonemes on a 2-D Space

Figure 3 illustrates the problems of cross-class errors and when viewed in the perspective of figure 2 it can be seen that many of the cross-class errors are confined to the fricatives, stops and affricates. If the place and manner of articulation of these phonemes is considered (Flanagan [1]; Levitt *et al.* [3]), the general findings here are not surprising.

5. INFORMATION TRANSMISSION

It was mentioned earlier that the two dimensional representation of phonemes lends itself quite naturally to the transmission of speech information via tactile means. For efficient information transfer, it is desirable to have the phonemes each occupying an equal area on the two dimensional space; or to have them organised such that the space occupied by each phoneme is proportional to its probability of occurrence. This gives a sparse arrangement for easier tactile detection, enhancing the sensitivity around the more common phonemes. The confusion picture of figure 2 has been non-linearly transformed to produce a first approximation of a regular grid arrangement for the purpose of information transmission and is shown in figure 4.

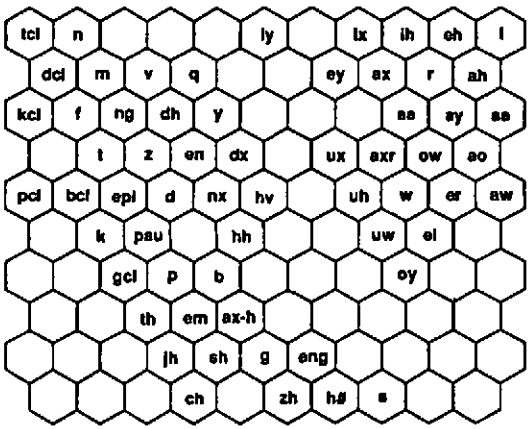

Figure 4: Regular Grid Arrangement of Phonemes

6. DEVELOPMENTS

To arrive at a better regular grid arrangement for the purpose of tactile communication or information transmission, the arrangement of figure 4 can be evaluated by applying the original data (in its probability vector form) and an attempt made to minimise the average variance of the phoneme excitation patterns that result. Adaptations of standard minimum variance criteria techniques (see Duda and Hart [9]) are then used to rearrange the elements of the grid to give the optimum phoneme arrangement.

To speed up the process of variance minimisation, the original mass of data shall be reduced using vector quantisation. Each probability vector in the original database is reduced to a finite set of quantised vectors effectively reducing the size of the database by a few orders. With just a limited set of probability vectors to apply to the regular grid arrangement the computing run time is substantially reduced. Since figure 4 is derived from the results of a proximity analysis based on phoneme confusions, it is not expected that the final arrangement will be much different.

7. CONCLUSIONS

A self-organising method of generating a low dimensional representation of the phonemes of the English language has been presented. The general findings are found to be consistent with work done by others but more importantly the confusion information of all the phoneme classes have been presented collectively enabling problems such as cross-class errors to be viewed more readily. It is hoped that by considering the loss in variance as a result of reducing the multi-dimensional phoneme space down to just two dimensions, a better understanding can be gained in the redundancy of current techniques used for speech recognition. The two dimensional format is also shown to have potential applications in tactile communication aids for the hearing impaired.

REFERENCES

[1] J.L. FLANAGAN, *Speech Analysis Synthesis and Perception*, Springer-Verlag, 1972.

[2] J.D. O'Conner, *Phonetics*, Penguin Books Ltd, 1986.

[3] H. LEVITT, J.M. PICKETT and R.A. HOUDE (Edited by), Prologue: *Sensory aids for the hearing impaired*, IEEE Press, New York, 1980.

[4] L.R. RABINER and R.W. SCHAFER, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[5] T. KOHONEN (1988). "The neural phonetic typewriter", IEEE Computer 21(3), pages 11-22.

[6] A.J. ROBINSON (1992). "A real-time recurrent error propagation network word recognition system", In *Proc.* ICASSP-92(3), pages 617-620.

[7] E. DAVID and P.B. DENES (Edited by), *Human Communication: A Unified View*, McGraw-Hill, 1972.

[8] P.E. GREEN, *Analyzing Multivariate Data*, The Dryden Press, 1978.

[9] R.O. DUDA and P.E. HART, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[10] L.F. Lamel, R.H. KASEL and S. SENEFF (1987). Speech Database Development: Design and Analysis of the Acoustic-phonetic Corpus. In *Proceedings of the DARPA Speech Recognition Workshop, pages 26-32.*

[11] J. HERTZ, A. KROGH and R.G. PALMER, *Introduction to the Theory of Neural Computation*, Addison-Wesley, 1991.