USING THE VOWEL TRIANGLE IN AUTOMATIC SPEECH RECOGNITION

David B. Grayden* and Michael S. Scordilis†

*Department of Electrical and Electronic Engineering
The University of Melbourne

†Department of Electrical and Computer Engineering
The University of Patras

ABSTRACT - An approach to reducing the number of insertions in a speech recognition system is presented which makes use of the relationship between the places of articulation of sonorant phonemes. A neural network is trained to locate place of articulation and the resulting contour is examined for phoneme boundaries. A Time-Delay neural network is also developed to locate nasal phonemes, a special case of sonorants.

INTRODUCTION

Phoneme detection and classification is an important first step in the recognition of large-vocabulary, speaker-independent, continuous speech. Any improvement in the phoneme recognition stage of a speech recognition system significantly improves lexical access, both in the accuracy of words hypothesised and in the amount of processing that must be done by the higher-level recognition stages.

Coarticulation between phonemes causes their properties to be context-dependent and it blurs their boundaries making them hard to locate. A key consideration in phoneme recognition is the development and utilisation of computational models that capture this acoustic variability of speech. While a variety of approaches exist, the use of detailed acoustic-phonetic knowledge in speech recognition is drawing considerable attention (Waibel & Lee, 1990).

At SST-94, a phoneme recognition system was presented that incorporated a hierarchy of neural networks and knowledge-based classification (Grayden & Scordilis, 1994). The system utilised acoustic knowledge for the analysis and representation of classes of phonemes. Moreover, phonetic knowledge was incorporated in the development of the structure of the recognition system itself. This approach facilitated the utilisation of knowledge about the properties of broad phonemic classes. This way, computation modules for recognition could be optimised for best performance in particular tasks. Intermediate results also allowed reliable broad category discrimination to be available. Such broad categorisation can provide significant reduction in lexical search during word hypothesis (Zue, 1985).

The majority of the errors made by the phoneme recognition system were in sonorant (1) segmentation and classification. Sonorants are highly variable across speakers and are susceptible to coarticulation effects. This makes them very difficult to accurately segment and classify.

This paper first of all reviews the performance of a Time-Delay Neural Network (TDNN) (Grayden & Scordilis, 1994), when performing sonorant segmentation and recognition. Then an approach to overcoming the problem of high insertion rate is presented. Knowledge of the vocal tract and the properties of sonorant phonemes are used to reduce the number of errors by tracing trajectories through the vowel triangle (Roach, 1992).

OVERVIEW OF PHONEME RECOGNITION SYSTEM

System Structure

A hierarchical phoneme recognition system was developed comprising modules best suited for processing particular phonemic groups (Grayden & Scordilis, 1994). Three basic acoustic characteristics of phonemes were used in the division of tasks: the obstruent/sonorant property, the manner of articulation and voicing, and the place of articulation classification.

The first classification module located regions according to two broad phonemic categories, sonorant and obstruent (1) (Edwards, 1992), which provided an initial classification of phoneme regions. Then, the manner of articulation classification divided these clusters into regions of finer phonemic distinctions. Finally, the place of articulation classification provided the recognition of individual phonemes. Both manner and place of articulation classification were achieved with neural network structures as well as rules extracted from detailed examination of relevant features in the regions of interest.

This paper will concentrate on sonorant segmentation and classification. Please see Grayden and Scordilis (1994) for a more comprehensive system overview and description of the other components of the system.

There are a number of manner of articulation classes within the sonorant group of phonemes. Vowels and diphthongs are phonemes which are uttered with a very open vocal tract. These sounds are usually the loudest and make up the nuclei of syllables (Roach, 1992). Diphthongs are transitory, with the place of articulation changing while the phoneme is uttered. The other sonorant phonemes (nasals, liquids and glides) have relatively more vocal tract constriction and are usually classified as consonants.

Corpus

The development and testing of the system was performed using the DARPA Acoustic-Phonetic Continuous Speech Corpus (TIMIT). This database contains a large number of fluently spoken sentences from 630 speakers of American English grouped into eight dialects, and it is divided into separate training and testing sets. Included with the corpus are time-aligned phonetic transcriptions. Speech samples in the TIMIT database are 16 bit integers at a sampling rate of 16 kHz. For testing the performance of individual modules and of the complete system, large numbers of phoneme exemplars or sentences were randomly selected from the test set. Only the 'SX' and 'SI' sentences were used in this work.

The first processing operation on the speech signal was the extraction of features from the utterances. The features were provided via a frame-based filterbank analysis. A 256 point Hamming window was applied at 5 ms intervals across the speech segment and the magnitude FFT was computed for each windowed speech portion. This was then compressed to a set of 16 mel-scale filterbank energies.

ALLSONORANTS TRAINING AND TESTING

Initially, a number of different hierarchies were investigated for sonorant manner of articulation classification. These involved breaking down the task in a number of different ways to arrive at the best configuration for sonorant recognition. Because of the high variability of sonorant phonemes, a single-pass approach proved superior to the others. A single Time-Delay neural network was trained for classification of all sonorant phonemes in the 24 phoneme sonorant set used by this system (Grayden & Scordilis, 1994). This was called the ALLSONORANTS TDNN.

Three layers of neurons, in addition to the input layer, were used. The input layer accepted signal feature vectors corresponding to fixed time intervals separated by 10 msec. These feature vectors were taken from the 16 mel-scale filterbanks energies by averaging pairs of adjacent mel-scale spectra, thereby halving the number of frames of data. An input size of 15 feature vectors was used, with the onset of each phoneme occurring in the third vector. The resulting input vectors were normalised to values between -1.0 and +1.0 as described in Grayden & Scordilis (1992).

The ALLSONORANTS TDNN was trained using 200 samples of each sonorant phoneme extracted randomly from the TIMIT training set. The training procedure was based on that described in Grayden & Scordilis (1992). For the training set, performance reached 55% . The resulting neural network was tested on 100 samples of each phoneme randomly extracted from the TIMIT test set; recognition performance in this case was 45%. The network performed best in recognising nasals and diphthongs while more often misclassifying vowels.

The ALLSONORANTS TDNN was used for both segmentation and recognition of the sonorant phonemes. The network was applied at 5 frame intervals within the sonorant regions. Where the winning output changed, a new sonorant phoneme was recognised.

With no segmentation of phonemes within sonorant regions, 23.5% of phoneme were deleted and 3.9% of boundaries were insertions. The insertions were due to the obstruent segmentation procedure. After incorporating the ALLSONORANTS TDNN, the deletions fell to 4.9%, but the number of insertions increased to 18.9%.

EXAMINATION OF RESULTS

The phoneme recognition system was tested over a large number of training sentences and a phoneme confusion matrix was constructed from these results. This allowed the performance of the system to be examined by showing which phonemes and phoneme categories were most often confused.

The confusion matrix showed that the manner of articulation task was performed quite well, especially in differentiating between obstruent and sonorant phonemes. Within sonorant phonemes, nasals were separated well from the remainder of the sonorants. However, liquids, glides, vowels and diphthongs were often confused.

Close examination of the sequence of phonemes produced by recognition of an utterance showed a pattern existed with most of the sonorant boundary insertions. When the phonemes were placed in their corresponding locations in the vowel triangle, it was observed that insertions were along the trajectory of transitions within the triangle. An example of this is shown in Figure 1 for the phoneme sequence /b ay z/ (buys). Instead of recognising one diphthong, /ay/, a sequence of phonemes were recognised along the path of the vowel triangle transition (/aa ay eh ih/). So the ALLSONORANTS TDNN was not really segmenting the speech into phonemes, but rather it tended to discriminate between vocal tract configurations.

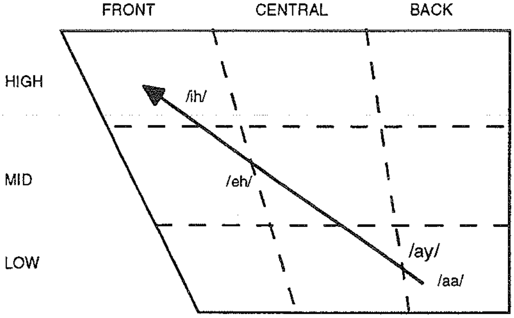

Figure 1. Vowel Triangle Example: Recognition of /ay/ in /b ay z/ (buys).

VOWEL TRIANGLE TRAINING AND TESTING

In order to compensate for this behaviour, neural networks trained to classify vowel triangle locations were investigated. In order to allow time-dependent vowel triangle location classification to be performed, a feed forward multilayer perceptron was used. The following architecture gave good performance:

Input layer:	10 feature vectors, each with 16 mel-scale filterbank outputs
First hidden layer:	72 neurons
Second hidden layer:	27 neurons

There were 9 output neurons, each corresponding to a region of the vowel triangle as shown in Figure 1.

The Vowel Triangle Network (VTN) was trained with a reduced set of vowels which nominally represented the different regions of the vowel triangle. These vowels are shown in Table 1. The training set was obtained from the TIMIT training set. The samples were extracted as 10 feature vectors, each separated by 10 msec, with the onset of the phoneme occurring in the third feature vector. The testing data was obtained in the same way from the TIMIT test set. The number of training and testing samples used were 500 and 100 of each phoneme, respectively. This created a training set of 4500 samples and a testing set of 900 examples.

VTN Output	Vowel Triangle Region	Phoneme
0	HIGH FRONT	/iy/
1	MID FRONT	/ey/
2	LOW FRONT	/ae/
3	HIGH CENTRAL	/er/
4	MID CENTRAL	/ax/
5	LOW CENTRAL	/ah/
6	HIGH BACK	/uw/
7	MID BACK	/ow/
8	LOW BACK	/aa/

Table 1. Vowel set used for training the Vowel Triangle Network.

The VTN was trained by backpropagation using the same speedups as the TDNN (Grayden & Scordilis, 1992). The network achieved a performance of 70% on the test set. The VTN was also tested with all vowel phonemes. A performance of 54% was achieved in locating them in the vowel triangle at their nominal locations.

VOWEL TRIANGLE SEGMENTATION

The vowel triangle was approximated as a square and the following formulae applied which effectively located the centre-of-gravity of the VTN outputs:

$$w_i = o_i + 1.0$$

$$t = \sum_{i=0}^{8} w_i$$

$$x = ((w_6 + w_7 + w_8) - (w_0 + w_1 + w_2))/t$$

$$y = ((w_0 + w_3 + w_6) - (w_2 + w_5 + w_8))/t$$

where o_i were the outputs of the neural network (values between [-1.0:1.0]), w_i were positive "weights" obtained from the outputs and *t* was the total weight. The final values *x* and *y* were coordinates *(x,y)* of the estimated place of articulation for the uttered sound. These coordinates were in the range [-1.0:1.0] where the origin (0,0) was the centre of the vowel triangle. Figure 2 shows a scatterplot of the coordinates of a number of instances of the vowels listed in Table 1. These were calculated from the outputs of the network when tested using 20 examples of each phoneme extracted from the TIMIT test set.

The information provided by the VTN was used to replace the segmentation function of the ALLSONORANTS TDNN by passing it over each sonorant region with a two frame step between classifications. The results were converted into *(x,y)* coordinates using the equations above. Then the "first derivative" was taken. This was a difference between coordinate values,

$$dx_i = x_{i+2} - x_{i-2}\ ,\ dy_i = y_{i+2} - y_{i-2}\ ,$$

and gave an indication of changing coordinates. The "second derivative" was also taken,

$$d2x_i = dx_{i+1} - dx_{i-1}\ ,\ d2y_i = dy_{i+1} - dy_{i-1}\ .$$

This was the important parameter which showed where changes were beginning and ending.

The next step was to locate peaks and troughs in the second derivatives. Any maximum or minimum values of magnitude greater than a fixed threshold (0.1) were marked. As they tended to indicate when transitions in the vowel triangle started or finished, these were possible locations of phoneme boundaries.

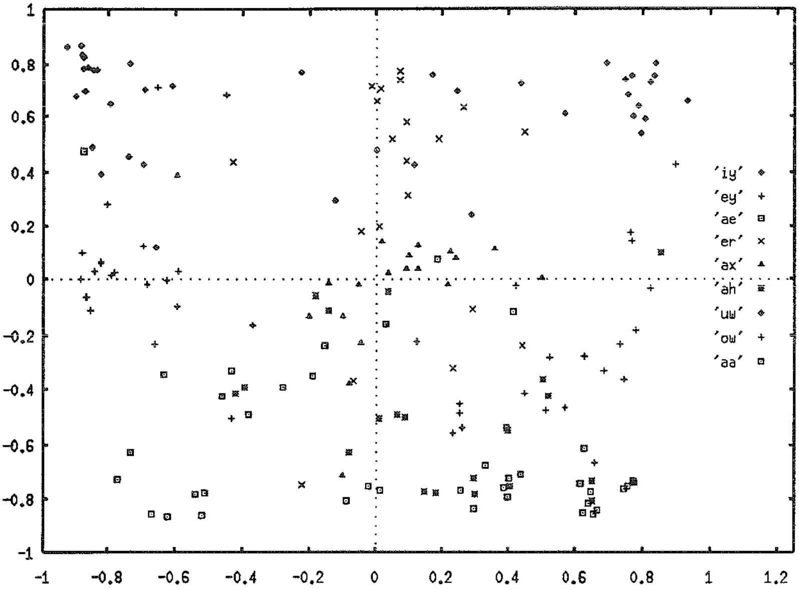

Figure 2. Scatterplot of vowel coordinates.

The marked locations (peaks) were then examined so that any not likely to indicate a boundary were removed. Rules were used to make these decisions. Firstly, the last peak in each sonorant region was removed if the *x* or *y* coordinate values rose or fell continuously for the rest of the segment. This was usually caused by coarticulation into a following obstruent phoneme outside the sonorant region.

The next rule removed peaks that did not correspond to significant changes in vowel triangle location. Often peaks existed where there were small rapid movements of the vowel triangle coordinates. These were not phoneme boundaries, but were most likely caused by inaccuracies in the VTN and the effects of voicing striations on the feature vectors. A peaks was removed if the Euclidean distance between its coordinates and those of the previous peak was less than a fixed threshold of 0.4.

Phoneme boundaries were placed at the locations of any remaining peaks and recognition was then performed at these boundaries.

SEGMENTATION OF NASAL PHONEMES

Nasals were not handled well by this technique. These phonemes have a significantly different manner of articulation since they involve complete closure in the mouth and opening of the nasal tract. There is some change in the vowel triangle coordinates since the mouth configuration changes, but some transitions, such as /iy/ to /n/ and /uw/ to /ng/ go unnoticed. In order to perform this segmentation, a TDNN was trained to discriminate between nasals and the rest of the sonorant phonemes. This network is referred to as the NASAL TDNN.

For training, 1000 instances of nasal phonemes and 1000 instances of other sonorant phonemes were extracted from the TIMIT training set. Ten feature vectors were extracted for each template with the onset of each phoneme in the third feature vector. There were two outputs.

After 460 training epochs, the recognition rate of the training set reached 96%. A test set was extracted from the TIMIT test set and this gave a performance of 94%.

The NASAL TDNN was applied to the system for segmentation before the VTN was used. NASAL was passed over the sonorant regions stepping 5 frames at a time and the output noted. Wherever the winning node changed, a boundary was placed.

RESULTS

The overall performance of the phoneme recognition system was tested on 100 randomly chosen sentences from the TIMIT test set. These sentences contained 3802 phonemes with approximately half being sonorant phonemes.

Using the ALLSONORANTS TDNN for segmentation of sonorant phonemes, 4.9% of phonemes were deleted and 18.9% of detected phonemes were insertions. When using the VTN TDNN, 7.6% of phonemes were deleted and 10.7% were insertions. After incorporation of the NASAL TDNN as well, the number of deletions fell to 6.9% while the insertions were 10.9%.

CONCLUSIONS

The phoneme recognition system that was presented in this paper was developed by incorporating knowledge about acoustics and phonetics in its structure. A further development in segmentation was presented which made use of the relationships between places of articulation of sonorant phonemes. A sonorant segmentation scheme using a neural network which located coordinates within the vowel triangle showed improved performance over segmentation using a Time-Delay neural network. A TDNN which located nasal phonemes also contributed to improving the performance of the system.

Phoneme recognition remained the task of the sonorant TDNN so significant improvements were not seen in that stage. However, the vowel triangle information will be valuable in the word hypothesis stage both in locating possible phoneme errors and reducing the lexical search space.

NOTES

(1) Sonorants include nasals, liquids, glides, vowels, and diphthongs. Obstruents include plosives, fricatives, affricates, and the glottal stop.

ACKNOWLEDGMENT

This work was supported by Telstra.

REFERENCES

Edwards, H.T. (1992) *Applied Phonetics: The Sounds of American English*, (Singular Publishing Group, San Diego).

Grayden, D.B. & Scordilis, M.S. (1992) "TDNN vs. Fully Interconnected Multilayer Perceptron: A Comparative Study on Phoneme Recognition", Fourth Aust. Int. Conf. on Speech Science and Technology, SST-92, 214-219.

Grayden, D.B. & Scordilis, M.S. (1994) "A Hierarchical Approach to Phoneme Recognition of Fluent Speech," Fifth Aust. Int. Conf. on Speech Science and Technology, SST-94, 473-478.

Roach, P. (1992) *Introduction to Phonetics*, (Penguin English, London).

Waibel, A. & Lee, K.-F. (Editors) (1990) *Readings in Speech Recognition*, (Morgan Kaufmann, San Mateo, California).

Zue, V.W. (1985) "The Use of Knowledge in Automatic Speech Recognition", Proc. IEEE 73, 1602-1615.