A HIERARCHICAL APPROACH TO PHONEME RECOGNITION OF FLUENT SPEECH

David B. Grayden and Michael S. Scordilis

Department of Electrical and Electronic Engineering
The University of Melbourne

ABSTRACT - An overview is presented of a hierarchical phoneme recognition system which performs the task in a number of steps: segmentation, manner of articulation classification and then place of articulation classification. A combination of knowledge-based techniques and neural networks are used within these modules.

INTRODUCTION

Automatic speech recognition (ASR) systems designed to handle multiple speakers, a large vocabulary and a large grammar must be able to recognise speech at the lowest level, typically the phonemic level, in order to achieve the required versatility. The phoneme recognition approach has proved popular for building many of the large vocabulary systems in existence (Waibel & Lee, 1990). Information derived from this basic level is used at higher levels of the recognition process such as word selection and syntax and semantic analyses (Lee, et. al., 1990, Sagayama, et. al., 1992, Deller, et. al., 1993).

Fluent speech is composed of sequences of words with little or no pauses between them. Since this is the natural form of speech for humans, it is preferred to uttering connected isolated words. However, when speaking fluently, the boundaries between words become difficult to locate as words blend into one another. Coarticulation is also more prevalent within words as well as between words. This leads to a sequence of phonemes that is different to that produced in isolated word speech, but human listeners are still able to understand without difficulty.

A hierarchical approach has been used in this system for recognition of phonemes in fluent speech. The tasks performed by the recogniser are divided into modules which relate to different acoustic and phonetic processes present in the utterances thus restricting the classification tasks that must be performed. The modular approach also allows flexibility in using different techniques to perform each task. Bayesian classifiers, neural networks and knowledge-based techniques are used where they can provide the best results. This method produces a versatile system that may be altered and adjusted as better classifiers are developed.

SYSTEM FRAMEWORK

An outline of the phoneme recognition system is shown in Figure 1. The first processing operation on the speech signal was the extraction of features relevant to the recognition process. In this system, the features were: 16 mel-scale filterbank energies, low and high frequency energies and a ratio of low frequency to high frequency energy.

Each sentence of speech was then segmented into phonemes with careful attention paid to accurately locating obstruent and sonorant regions. This was performed using a combination of Bayesian classifiers and Time-Delay Neural Networks (TDNNs) (Grayden & Scordilis, 1994).

The next stage in the system classified the phonemes into manner of articulation classes and, finally, determined the place of articulation of each phoneme. Both of these stages were performed by TDNNs (Grayden & Scordilis, 1993).

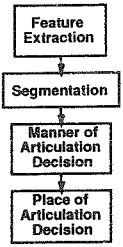

Figure 1: System Framework

SPEECH DATA

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) was used for all system training and testing. Only the 'SX' and 'SI' sentences have been used in this work. Table 1 shows the manner of articulation classes used by the system and the phonemes within these classes (labelled as provided by TIMIT). The phonemes /q/ (glottal stop), /dx/ (flap) and /v/ were also classified as "dip" phonemes as discussed below.

Silence & Closures	/h#/ /pau/ /epi/ /bcl/ /dcl/ /gcl/ /pcl/ /tcl/ /kcl/
Voiced Plosives	/b/ /d/ /g/
Unvoiced Plosives	/p/ /t/ /k/
Affricates	/jh/ ch/
Voiced Fricatives	/z/ /zh/ /v/ /dh/
Unvoiced Fricatives	/s/ /sh/ /f/ /th/ /hh/
Nasals	/m/ /n/ /ng/ /em/ /en/ /eng/ /nx/
Liquids	/l/ /el/ /r/
Glides	/w/ /y/
Vowels	/iy/ /ih/ /ix/ /ey/ /eh/ /ae/ /aa/ /ao/ /ow/ /uh/ /uw/ /ux/ /er/ /ax/ /ax-h/ /axr/ /ah/
Diphthongs	/aw/ /ay/ /oy/

Table 1: Manner of Articulation Classes

The phonetic labeling in TIMIT differentiates between the closure and release portions of plosive and affricate phonemes and classifies the closure regions as /bcl/, /dcl/, /gcl/, /pcl/, /tcl/ and /kcl/ depending on the voicing condition and the location of the closure in the vocal tract. The developed system was only required to classify the closures as voiced or unvoiced. TIMIT also contains three categories of silence: /h#/ (beginning and ending silence), /pau/ (pause) and /epi/ (epenthetic silence). These were all combined into one silence category in this work.

Among the nasals, the following groups of phonemes were also combined into single phonemes: /m/ and /em/; /n/, /en/ and /nx/; /ng/ and /eng/. The liquids /l/ and /el/ were combined as were the glottal fricatives /hh/ and /hv/. The following pairs of vowels were also combined: /ih/ and /ix/; /uw/ and /ux/; and /ax/ and /ax-h/.

FEATURES

The Time-Delay Neural Networks (TDNNs) required 16 mel-scale filterbank energies for each frame (Grayden & Scordilis, 1992). These were extracted from TIMIT data by applying a Hamming window to a 256-point frame of data, taking the FFT and then scaling to mel-scale.

This window was shifted by 80 samples for each frame of features providing feature vectors which were 5 msec apart. For input to the TDNNs, adjacent pairs of vectors were then combined to give 10 msec separation and normalised to $[-1 : +1]$.

From the mel-scale filterbank energies, further features were extracted for use by the segmentation algorithm (Grayden & Scordilis, 1994). These are shown in Table 2.

LOW	0–1000 Hz frequency band energy
HIGH	1000–8000 Hz frequency band energy
DIV	$\frac{0-300\text{Hz}}{3700-7000\text{Hz}}$ frequency band energy ratio
DISTANCE	Euclidean distance between mel-scale vectors

Table 2: Manner of Articulation Classes

SEGMENTATION

Segmentation was performed by making use of the acoustic features of phonemes. The segmentation system framework is illustrated in Figure 2. Initially, obstruent and sonorant regions were discriminated using the LOW and DIV features. This task was called "SPFA vs. ELSE" discrimination where SPFA was made up of Silence, Plosives, Fricatives and Affricates, and ELSE was all the remaining phonemes. From a large number of training sentences, a decision surface was constructed that could be used to classify further sentences. High performance was achieved in this stage, with most of the errors occurring for short, high-energy obstruents (such as /q/, /dx/, /v/) when they occurred between sonorants. When not including transient frames between phonemes, the frame error rate was around 5% (Grayden & Scordilis, 1994).

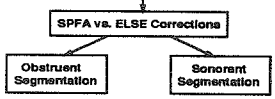

Figure 2: Segmentation System Outline

The next step in segmentation was to correct the errors made by SPFA vs. ELSE. The short obstruents were often clearly evident in the LOW feature as short dips in energy resulting from brief obstructions in the vocal tract. Thus, the derivative of LOW was taken and examined for these short events. This algorithm located almost all instances of these phonemes with only 2.5% insertions. The insertions were usually nasals and short liquids.

After locating SPFA and ELSE regions, further segmentation between individual phonemes was performed. Since the obstruent and sonorant regions were separated, different techniques could be used to segment within each region. Obstruents have markedly different vocal tract configurations and manners of articulation so the SPFA phoneme boundaries could be located by examining LOW and HIGH for any abrupt or large changes. Using this method, around 85% of obstruent boundaries were accurately located with 8% insertions. Many of the deletions were due to the inability of the algorithm to locate very short events such as plosive releases.

Sonorant phonemes are very similar to each other so the locating of boundaries between them was much more difficult. The DISTANCE measure was first examined to locate regions where the spectrum changed significantly. This provided the locations of the more obvious sonorant-sonorant boundaries although some insertions within diphthongs were observed.

The regions between located sonorant phonemes were then examined to locate more subtle sonorant-sonorant boundaries. Initially, a TDNN that distinguished between nasals, liquids, glides, vowels and diphthongs (NLGVD) was passed over the features. TDNNs were designed with the property of being somewhat time-invariant (Waibel, et. al., 1989, Grayden & Scordilis, 1992) so that as the network was passed over the speech, the output indicated the manner of articulation of the phoneme being examined. When the output changed, a boundary was placed. This network examined only 60 msec of speech in order to reduce overlap between boundaries. Then, a TDNN recognising all 24 sonorants was applied within the located NLGVD regions in order to locate any boundaries between phonemes with the same manner of articulation. The manner of articulation network was used before the all-sonorants network as its performance was higher for the broader manner of articulation classes. The all-sonorants network also required 15 frames of input data as its performance decreased markedly when fewer input frames were used. Most boundaries could be located in this manner but, owing to the imperfect time-invariance of the TDNN, a high number of insertions also resulted.

When examining 500 test sentences, the overall segmentation system located 88% of the phoneme boundaries with 1.4% SPFA vs. ELSE classification errors and an extra 22% of inserted boundaries.

MANNER OF ARTICULATION CLASSIFICATION

Classification of the manner of articulation of obstruent phonemes was performed using TDNNs. Firstly, a neural network that distinguished between silence, plosives, fricatives and affricates (S/P/F/A) was applied at the phoneme boundary locations determined by the segmentation module. This TDNN could achieve 87% accuracy on TIMIT testing data. The results from the network were then arranged in descending order of probability. This was done using the formula

$$\Pr(M_i) = \frac{y_i + 1}{\sum(y_j + 1)} \tag{1}$$

where M_i was manner of articulation class i and y_i was the value of output neuron i. $\sum(y_j + 1)$ was the total of all the TDNN output values scaled to $[0 : 2]$ as all TDNNs in this work had outputs in the range $[-1 : +1]$ (Grayden & Scordilis, 1992).

The next step was to determine whether the phoneme was voiced or unvoiced. This was done using a number of TDNNs. Different networks were used for plosives, fricatives and affricates giving results of 75%, 86% and 87% respectively on TIMIT testing data. It was found that the TDNNs could provide good voiced/unvoiced decisions provided the manner of articulation was restricted. Future work will incorporate other techniques to provide more accurate classification.

The task of classifying sonorants into manner of articulation categories was performed by the segmentation algorithm. Nasals, liquids, glides, vowels and diphthongs were separated using a TDNN passed over the sonorant regions. The network outputs were then arranged according to probability in the same way as the obstruents.

PLACE OF ARTICULATION CLASSIFICATION

Obstruent phonemes were classified using further TDNNs (Grayden & Scordilis, 1993). Each phoneme instance was applied to separate neural networks that distinguished between b/d/g, p/t/k, z/zh/v/dh, s/sh/f/th/hh and silence types. The silence type network distinguished between silence, voiced closures and unvoiced closures, while the other networks decided between the phonemes as shown. The affricate phonemes were not included here as this decision was already made by the voiced/unvoiced decision. Recognition probabilities were determined from network outputs and multiplied to the results of S/P/F/A and, if necessary, the results of voiced/unvoiced decisions. Finally, all the phonemes were combined into a single list in decreasing order of probability and the top phoneme chosen.

The obstruents detected by dips in low frequency energy within sonorant regions were treated as a special case when classifying the phonemes. Since the phonemes that appeared in these locations were only a very small subset of the obstruent phonemes, they were classified using a TDNN that distinguished between glottal stops (/q/), flaps (/dx/), /v/ and nasals.

The classification of sonorant phonemes was found to be best performed by a one large neural network. A TDNN was trained to distinguish 24 sonorant phonemes as outlined for the segmentation system. As for obstruents, a list of phonemes is provided with associated probabilities.

PERFORMANCE

Testing was performed using 500 'SX' and 'SI' sentences randomly chosen from the TIMIT testing set. No restrictions were made regarding dialect regions or gender of speakers.

The criterion used for determining correct segmentation was that phonemes located within four frames (20 msec) of the location indicated by the TIMIT phoneme transcription were correctly positioned. This width was chosen since the TDNNs were found to be most time-invariant within this region (Grayden & Scordilis, 1992, Grayden & Scordilis, 1993). Over the 500 test sentences, there were 12.1% deletions while 21.7% of phonemes detected were insertions. These results inevitably include some deletion-insertion pairs due to incorrect positioning of phoneme boundary locations.

The segmentation stage also produced some errors in SPFA vs. ELSE decisions. The errors in distinguishing between obstruent and sonorant regions showed in 1.4% of the detected phonemes. Most of these errors occured for short liquids that were mistaken for 'dip' phonemes.

Of the correctly detected phonemes, there were 21.1% manner of articulation classification errors and a further 21.5% place of articulation errors. This produced a recognition performance of 56% for correctly segmented phonemes and 65% when considering the top two phoneme choices.

Overall, 44% of the phonemes in the TIMIT test sentences were located AND correctly classified. This increased to 53% for the top two choices. Among the 500 test sentences, the best result was 70% and the worst was 16%.

DISCUSSION

The number of insertions can be significantly reduced by enforcing tighter restrictions on the sonorant segmentation performed by the TDNNs but at the expense of more deletions. It is believed that the following task, that of word hypothesis, would benefit more by having fewer deletions than fewer insertions.

Manner of articulation classification can be improved by a more sophisticated voiced/unvoiced decision algorithm. The use of classical techniques and other neural networks should increase this performance considerably. The use of more neural networks to classify manner of articulation of sonorant phonemes may provide more information to the word hypothesis system and attempt to take into account the large variability in duration that is a characteristic of sonorant phonemes (Edwards, 1992, Hataoka & Waibel, 1990). Place of articulation classification will also benefit by further investigation of other TDNN architectures to overcome the problems of sonorant duration variabilities.

The phoneme recognition system presented provides important manner of articulation information that will be very useful in the word hypothesis stage. The size of the vocabulary that must be searched is reduced considerably (Zue, 1985). When the SPFA vs. ELSE regions of a word are known, a lexical search of the 100,000 word CMU public domain pronunciation dictionary

(cmudict.0.1) can be reduced to a search of 4000 words on average. When manners of articulation are correctly classified, the search reduces to an average of 3 words. This indicates that the phoneme recognition system developed can provide well for a word selection stage.

ACKNOWLEDGMENT

This work was supported by Telecom Australia.

REFERENCES

Deller, J.R., Proakis, J.G., Hansen, J.H.L. (1993) *Discrete-Time Processing of Speech Signals,* (Macmillan Publishing Company: New York).

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", Proceedings of the Fourth Australian International Conference on Speech Science and Technology, SST-92, pp. 214–9.

Grayden, D.B., Scordilis, M.S. (1993) "Recognition of Obstruent Phonemes in Speaker-Independent Fluent Speech Using a Hierarchical Approach", Proceedings of the Third European Conference on Speech Communication and Technology, Eurospeech '93, pp. 855–8.

Grayden, D.B., Scordilis, M.S. (1994) "Phonemic Segmentation of Fluent Speech", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94, pp. I–73–6.

Hataoka, N., Waibel, A.H. (1990) "Speaker-Independent Phoneme Recognition on TIMIT Database Using Integrated Time-Delay Neural Networks (TDNNs)", Proceedings of the IEEE International Joint Conference on Neural Networks, IJCNN-90, pp. I–57–62.

Lee, K.F., Hon, H.W. Reddy, R. (1990) "An Overview of the SPHINX Speech Recognition System", IEEE Trans. Acoust. Speech Signal Process., Reprinted in (Waibel & Lee, 1990).

Sagayama, S., et. al. (1992) "ATREUS: Continuous Speech Recognition Systems at ATR Interpreting Telephony Research Laboratories", Proceedings of the Fourth Australian International Conference on Speech Science and Technology, SST-92, pp. 324–9.

Waibel, A., Hanazawa, T, Hinton, G., Shikano, K. & Lang, K. (1989) "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Transactions on Acoustics, Speech and Signal Processing, 37(3), pp. 328-39.

Waibel, A. Lee, K.F. (Editors) (1990), *Readings in Speech Recognition*, (Morgan Kaufmann: Palo Alto).

Zue, V. (1985) "The Use of Speech Knowledge in Automatic Speech Recognition", Proceedings of the IEEE, Vol. 73, pp. 1602–15.