EXPLORING THE PHONETIC STRUCTURE OF THE SPEECH SIGNAL USING MULTI-LAYER PERCEPTRONS

Shuping Ran and J.Bruce Millar

Computer Sciences Laboratory
Research School of Physical Sciences
Australian National University

ABSTRACT - Two experiments using a multi-layer perceptron to explore phonetically significant boundaries in the speech signal are described. The two fundamental distinctions, between speech segments which have periodic or aperiodic waveforms, and between speech segments which have transitional or steady state spectra, are examined to lay the foundation for possible future work. In the first experiment the refinement of hand segmented vocalic nuclei is shown to be possible for at least one speaker, whereas in the second experiment new boundaries are created using criteria developed from within the data itself.

INTRODUCTION

Speech is a continuous signal generated by moving the articulators through a coordinated series of gestures guided by the target articulation of a sequence of phonemes (Potter et al, 1947; Fant 1960; Flanagan 1972). Such movement blurs the boundaries between the realizations of adjacent phonemes which, in their complexity, admitt a large amount of variability. In order to relate acoustic processing to phonetic structure, it is necessary to derive a time alignment of phonetic descriptions of the speech to the acoustic signal.

Many different techniques are used to perform this task in the field of automatic speech recognition. Some of these techniques require segmentation of the speech into sub-word units such as allophones, phonemes, diphones, or syllables. These units may be identified either on a linguistically motivated basis such that they correspond roughly to perceived phonetic segments, or strictly on the basis of acoustic homogeneity (Zue 1980). Traditionally, this segmentation has been done manually by a experienced person, by visual inspection examining the waveform of the acoustic signal with the aid of graphic displays of the energy contour or spectrogram. However, this process is extremely tedious and time consuming. The decisions are subject to human errors such as mechanical errors committed by hands or mis-interpretation of a spectral or waveform display. An automatic method is therefore preferable.

There are two main approaches for the automatic segmentation of the acoustic signal. One of them transfers segmentation data from an utterance of identical content, which has already been segmented, onto the utterance requiring segmentation. This method uses a reference waveform which may be hand segmented and labelled natural speech of a reference speaker (Wagner 1981), or may be synthetically generated utterances from a known phonetic string, in which case, no manual segmenting and labelling is needed (Bridle and Chamberlain 1983). Unless a sufficiently high-performance text-to-speech system can generate the reference utterance the task of hand- segmentation and labelling is still required for the generation of reference material. The other approach does not have to compare with a reference waveform. One example, given by Wilpon (1987), is based on the variation of the spectral contour. The speech is segmented automatically into sub-word units which are defined acoustically, but not necessarily phonetically. The lack of the phonetic interpretation of the segments makes it hard to incorporate such segmented data with recognition procedures if only the phonemic transcription of the word is known.

There is scope therefore for the development of methods to automatically segment speech signals in a manner that is sensitive to their phonetic structure. From an acoustic-phonetic point of view, the speech signal consists of two major types of acoustic signal, periodic and aperiodic. The periodic signal repeats itself every T seconds, where T is called the period of the waveform (Lynn 1980). For all the periodic sounds that occur in the course of speech, the sound source is the larynx. Periodicity is present in all vowel sounds and voiced consonants. The aperiodic signal has an irregular form and

is generated by a turbulent air-stream which is heard as noise. Aperiodicity is present briefly in plosive consonant sounds and on a more continuous basis in the fricatives and affricates (Fry 1980). The periodic - aperiodic distinction is fundamental to the acoustic phonetic structure, thus it is seen as the appropriate starting point for a phonetically motivated segmentation system.

Another fundamental distinction in acoustic phonetics is that between transitional and steady-state sounds. As the acoustic realization of a phoneme depends on the immediate phonemic environment - the so-called coarticulation phenomenon - there is often a rather rapid change in the spectral structure of the speech signal in the vicinity of phoneme boundaries. At other times the spectral structure is relatively constant. This effect can be easily observed in a spectrogram (Potter et al, 1947).

In this paper, a method is presented which allows a rough segmentation to be developed manually which separates the signal into its periodic and aperiodic parts, and then automatically refines this segmentation. This is followed by a method which separates the transitional parts and the steady parts within a periodic signal segment.

SPEECH DATA CORPUS AND ANALYSIS

The experiments were conducted on five repetitions of CVd syllables from one native Australian English speaker, where C represents the voiced and voiceless stop consonants: [p t k b d g]; V represents the eleven Australian nominally monophthongal vowels : [i ɪ ɛ æ ɑ ʌ ɒ ɔ ʊ u ɜ]. The acoustic signal was sampled at 10kilosamples/second, and split into a sequence of 12.8ms time frames which had 6.4ms overlaps. The frames were then Hamming-windowed.

The speech analysis consisted of the calculation of 10 Mel-scale Frequency Cepstral Coefficients (MFCC) at 6.4ms intervals. The MFCC were derived using the algorithm presented by Davis and Mermelstein (1980). It should be noted that the MFCC is not ideally suited to represent periodicity in general, as its sensitivity to harmonic structure is limited. However in the present restricted data corpus periodicity has been interpreted as that signal within the limits of the vocalic nucleus. This is strictly true for the unvoiced consonant onsets, but has the effect of classifying the pre-voicing of the voiced stop onset as aperiodic. In a more general data corpus a more appropriate pre-processing of the signal could be used.

THE MULTI-LAYER PERCEPTRON

The method developed in this study relies upon the characteristics of the multi-layer perceptron (MLP). In the connectionist approach to pattern classification, reference data are represented as multiplicative weights linking the nodes of a network in which each node is a simple processing unit (Lippman, 1987; 1988). In the following experiments a fully - connected MLP of three layers - one input layer, one hidden layer and one output layer was used. The number of nodes per layer used in each experiment described in the following sections was different. Figure 1 represents a example of a three layer (MLP), where the input layer has ten nodes, the hidden layer has four nodes and the output layer has two nodes.

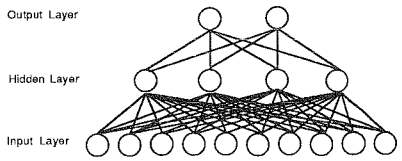


Figure 1. A three layer Multi-Layer Perceptron.

The input to the MLP is a group of patterns which consist of feature vectors, in this study, the 10 MFCCs. The output of the MLP consists of N nodes where N is equal to the number of classes into which the input patterns are to be classified. Each output node corresponds to one class.

It is necessary to distinguish clearly between two modes of operation for the MLP: Training mode and recognition mode. In training mode, training data is presented to the input nodes of the MLP while

the desired output is compared with the values on the output nodes. The global error between the desired and observed output is used to determine changes to the weights. These changes are applied after each complete presentation of the training data. The changes to the weights are determined by the conjugate-gradient optimization algorithm (Powell 1977). The training process terminates when the global error meets the minimum acceptable value (given by the experimenter), or when it could not find a better way to minimize the global error, which is the case when it encounters a local minimum of the error space.

In recognition mode, one pattern is presented at the input at one time. The output nodes compete to represent the input pattern. The node with the highest activation score wins the process, and the input pattern is classified to the class represented by this output node.

A critical characteristic of the MLP for the purpose of this study is that there is a positive correlation between the number of hidden nodes that are used, and the complexity of the decision boundaries which the network can encode in its weights (Lippman 1987). For a given problem, the number of hidden nodes to be used for the MLP is related to the grade of detail which it is desired to encode to represent the relationship between the input pattern space and the categories of the output space. With a smaller number of the hidden nodes, the MLP could encode only more general relationships, and with a larger number of hidden nodes the MLP could encode more detailed relationships.

PERIODIC DISTINCTION

The first experiment used a MLP to segment the CVd signals into periodic and aperiodic parts, where, as noted above, the periodic parts correspond to the vocalic nuclei. This experiment was motivated by an other experiment, where the same data was used, and in which it was intended to recognise the six stop consonants and the eleven Australian monophthongs. The data was segmented by hand, and each CVd syllable was segmented into the vocalic nucleus and the remaining part of the syllable. The initial recognition result based on frame by frame recognition was relatively poor. One of the analyses to investigate the reason for this poor performance was to examine which part of the time course of the utterance failed to be recognised correctly. A large number of such failures were found at the boundaries of the segmentation between the vocalic nuclei and the remaining part of the syllable. The initial hand-segmentation of the data was re-examined and it was found that those failures corresponded to errors in the hand-segmentation. After correcting these errors of hand-segmentation, the recognition experiment was repeated, and the recognition result improved substantially. In this way, a new method of refining a rough hand segmentation using a MLP was discovered.

THE METHOD

The method for refining rough segmentation operated as follows. The MLP was trained to criterion using the roughly segmented and labelled data. The data was hand-labelled by deciding on a segmentation point between the initial consonantal release and the vocalic nucleus of the syllable and then labelling each individual frame on either side of the segmentation point respectively as aperiodic or periodic. The trained MLP was then used in recognition mode on its own training data on a frame by frame basis. Frames whose previous labelling was not in agreement with the MLP output for that frame were noted. Where these frames occurred adjacent to a segmentation point, that is, where the previous labels changed from aperiodic to periodic or from periodic to aperiodic, the output of the MLP was deemed to be correct and the segmentation point was shifted to minimise the error in that region.

The MLP was then re-trained using the new segmentation point. The process was repeated until errors in recognition of the frames labelled periodic or aperiodic frames was reduced to zero or to an agreed level.

THE ARCHITECTURE

The architecture of the MLP used was a 10 node input layer, a 4 node hidden layer and a 2 node output layer. The inputs were the 10 Mel-scale Frequency Cepstral Coefficients, and the two output nodes corresponded to the two output classes: periodic and aperiodic.

THE RESULT AND EVALUATION

Using this method, on a rough segmentation of the five data sets, each containing one repetition of the 66 CVd utterances, the maximum number of frames of error corrected was 9 frames, corresponding to

approximately 58ms. This occurred on the item / bi / in which the segmentation point had originally been erroneously placed at the onset of pre-voicing. The new position was immediately before the nucleus onset.

In order to evaluate this method, the segmentation points of all utterances were intentionally shifted, randomly to the left or right, from the previously refined positions by a random number of frames. The intentionally perturbed data was then used in place of the rough hand-segmentation to test the robustness of the method. The results of this analysis is shown in Figure 2.

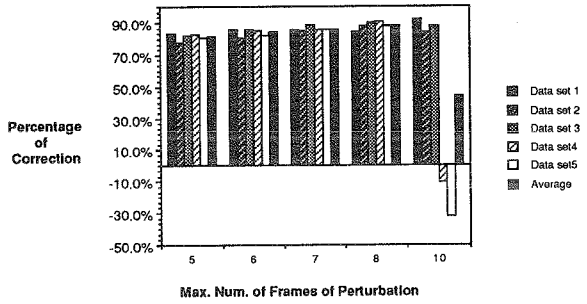


Figure 2. Result of analysis for intentionally perturbed data.

VOWEL NUCLEUS ANALYSIS

From the previous experiment, the rough segmentation was finally refined so as to determine accurate endpoints for the vocalic nucleus. From phonetic knowledge, we know that the stop consonants of English [p t k b d g] share the same manner of articulation, in that they are produced with the rapid release of a complete closure in the vocal tract (Lieberman and Blumstein 1988). There are several acoustic cues that contribute to the perception of a stop consonant, one of them is the formant transition into or out of the adjacent vowel (Lieberman and Blumstein, 1988; Fry, 1980). So the transition part of the vocalic nucleus for a vowel in different contexts (with different stop consonants) is different, even for the same vowel. Our aim was to separate the transition part of the nucleus from the steady part, in order to prepare the vocalic nucleus for further analysis.

The separation of steady-state portions of the vowel nucleus from its transitional portions is based on the fact that although the whole vocalic nucleus is initially labelled with a single vowel label, its transitional parts are, on a frame by frame basis, more similar to other vowels. It is assumed that the initial training of the MLP over a wide range of consonant contexts encodes a robust model of the vowel target despite the contradictory information contributed by frames in the consonantal margins of the nucleus. The specificity of the encoded vowel models will depend on a number of factors such as the rate of articulation of the syllable, the complexity of the MLP, and the composition of the training data. In this study there were eleven vowels in six different contexts for each repetition. The steady part of each of the eleven vowels was repeated six times, while the transition part of the vowel for each of the sixty-six syllables differ from each other. The accuracy with which the boundaries between steady states and transitions can be located will also depend on the local density of vowels in the vowel space as the definition of a transitional frame is related to its proximity to a neighbouring vowel. However this method allows the automatic segmentation of a vocalic nucleus so that transitional segments can be re- labelled with the joint labels of their adjoining segments and modelled at a further stage in an appropriate manner.

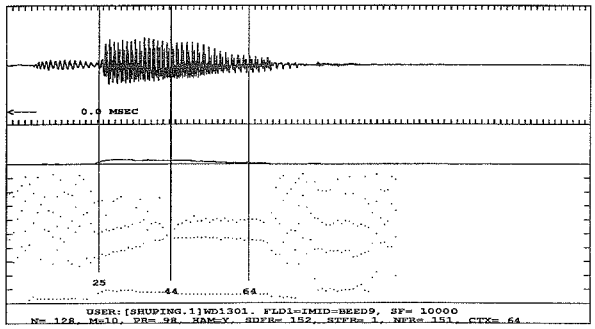
ARCHITECTURE

The choice of the number of hidden nodes which enable training of the network just to learn to classify the steady part of the vowels but not the frames in the transitional regions is clearly important owing to the correlation between number of hidden nodes and boundary complexity noted above. For this

experiment 6, 10, and 23 hidden nodes were evaluated. However only minor difference in the performance of the network was found when discriminating between transitions and steady state of the nucleus. The main experiment was based on a MLP with an input layer of 10 nodes, a hidden layer of 10 nodes, and an output layer of 11 nodes.

METHOD

Accurately delimited vocalic nuclei of the utterances were labelled on a frame by frame basis using the eleven classes of monophthongs as the labels. The MLP was trained using these labelled data, then tested using the training material. The activation scores of each of the output nodes for each frame were recorded. It was observed that the frames which had high activation scores in the correct output node corresponded to the steady part of the vowel. The frames which corresponded to the transition part of the vowel had their highest activation scores in output nodes which corresponded to different vowel labels. In this way, it was possible to separate the transition part from the corresponding steady part of the vocalic nucleus.



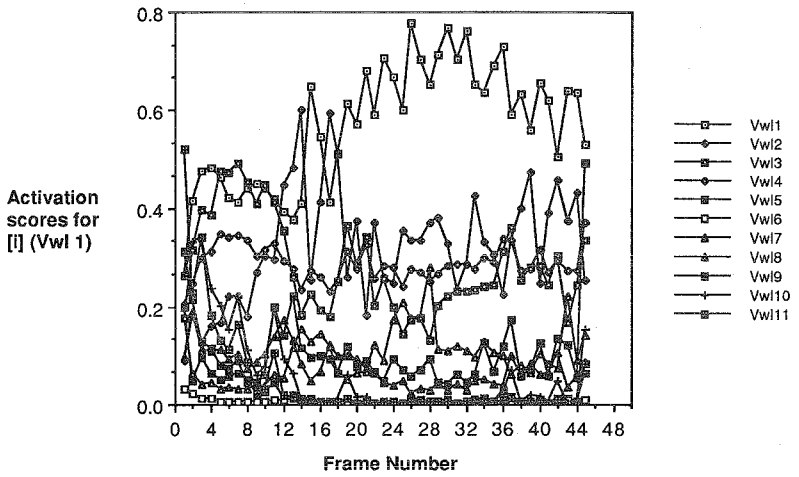


Figure 3.a. Waveform, energy, spectrogram for the word "Beed ";
3.b. Activation scores of the MLP for the vowel [i].

The result of this analysis is shown in Figure 3, where 3.a shows the activation scores of the MLP

in the recognition mode, and 3.b shows the the waveform, energy contour and the spectrogram. The beginning of the vocalic nucleus indicated as frame 1 in 3.a, corresponds to the 25th frame in 3.b.

DISCUSSION

These procedures are specific examples of a general process in which a limited number of category judgements are made about each frame comprising the data. The boundaries between these categories, as measured in the input-space, are encoded in an MLP of limited complexity. This encoding will capture the dominant features in the training data that correlate with the category differences. The positions of the category boundaries in the time domain are then successively refined using the errors generated between categories and the assumption of 'temporal continuity of category'.

Note that a more complex MLP may be able of encode the errors on which it is trained as legitimate alternatives of the truth. Such an MLP would not be useful in this method. The question of the limits on the complexity of the MLP to make it useful for this task has not yet been fully explored.

There are a number of parameters which would appear to be critical in this approach: first, the definition of the input space appropriate for the drawing of boundaries between the desired categories; second, the architecture of the MLP selected, and third, the degree of training given to the MLP. All of these need investigating before the method can be generally applied to the exploration of the phonetic category space of the acoustic signal of speech.

ACKNOWLEDGEMENT

The authors achnowledge the work of F. Clermont in collecting, digitising and organising the CVd data used in this study.

REFERENCES

Bridle J. S. and Chamberlain R. M. (1983). *Automatic Labelling of Speech using Synthesis-by-Rule and Non Linear Time-Alignment*, Speech Communication 2 (1983), pp187-189.

Davis, S. B. and Mermelstein P. (1980). *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Acoustics, Speech, and Signal Processing, (ASSP) Vol. 28, No. 4, (August 1980) pp357-366.

Fant, G. (1960). *Acoustic Theory of Speech Production*, Mounton and Co.,'s- Gravenhage, Netherlands, 1960.

Flanagan, J. L. (1972). *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.

Fry, D. B. (1980). *The Physics of Speech*, (Combridge University Press).

Lieberman, P. and Blumstein, S. E. (1988). *Speech Physiology, Speech Perception, and acoustic phonetics*, (Cambridge University Press), 1988.

Lippman, R. P. (1987). *An Introduction to Computing with Neural Nets*, IEEE ASSP Vol 4, No. 2, pp.4-22.

Lippman, R. P. (1988). *Neural Nets for Computing*, IEEE ICASSP'88 pp.1-6, New York, April 1988.

Lynn, P. A. (1980). *An Introduction to the Analysis and Processing of Signals* , (The Macmillan Press Ltd, Hong Kong).

Mermelstein P. (1975). *Automatic Segmentation of Speech into Syllabic Units*, J. Acoust. Soc. Am, Vol. 58, No. 4, October 1975, pp880-883.

Potter R. K., Kopp, G. A., and Green H. C. (1947). *Visible Speech*, (New York: Van Nostrand. Co.).

Powell, M. J. D. (1977). *Restart Procedures for the Conjugate Gradient Method* , Mathematical Programming, Vol. 12, pp.241-254, April 1977.

Wagner W. (1981). *Automatic Labelling of Continuous Speech with a Given Phonetic Transcription using Dynamic Programming Algorithm*, IEEE Conf. on Acoustics, Speech and Signal Processing, pp1156-1159.

Wilpon, J. G., Juang, B. H., Rabiner, L. R. (1987). *An investigation of the Use of Acoustic Sub-Word Units for Automatic Speech Recognition*, IEEE Int. Conf. Acoust. Signal Process., pp.821-824.

Zue, V. W. (1980). *Acoustic Processing and Phonetic Analysis* in Trends in Speech Recognition, edited by Wayne A. Lea (Prentice-Hall Inc.) chap. 5.