PHONETIC FEATURE EXTRACTION USING ARTIFICIAL NEURAL NETWORKS

Shuping Ran and J.Bruce Millar

Computer Sciences Laboratory
Research School of Physical Sciences and Engineering
Australian National University

ABSTRACT - The work described in this paper is part of a strategy to investigate useful architectures of parallel computation which encode speech knowledge into a speech recognition system to optimise its performance. At the first level of the system, phonetic features (an extended set of Jakobson et al's distinctive features) are extracted from burst onset intervals and pseudo-static vowel intervals of CVd words. Our result gives limited support to the existence of invariant cues for some of these features.

INTRODUCTION

After 40 years of research the topic of speech recognition has yielded two main streams of approach. On the one hand there is the knowledge-based pattern recognition approach in which general rules governing the structure of speech are used to interpret the speech data and to classify the inherent speech segments. On the other hand there is the data-driven pattern-matching or statistical approach in which classification transforms or statistical models of the data are built from a set of training data and then used to classify further speech data that were not a part of the training data.

The knowledge base for rule-based recognition has been accumulated over many decades of study in phonetic and linguistic science, and other related areas. Valid techniques for the training of data-driven approaches have been developed within the last decade and have achieved high levels of performance. This indicates that there exists structure within speech data in the form of phonetically relevant cues which are not adequately represented in the rule-based models. However the general tendency to use data-driven systems which have an "ab initio" approach to learning their patterns appears to be an over reaction.

The need to use speech knowledge in automatic speech recognition has been highlighted in the literature. Zue (1985) concluded that the reason for the slow advance in the automatic recognition of speech is due to the abandonment of phonetically based approaches in favour of the general pattern matching technique. Huckvale (1990) proposed to encode the speech knowledge into the system's structure.

Within a general paradigm of investigating architectures of parallel computation that may be optimised for the recognition of speech, we have been investigating ways of encoding speech knowledge into the system to enable better performance in the presence of the major "noise factors" for any model of speech: speaker variability and the effect of coarticulation. It is also hoped that this paradigm of enriching the system with speech knowledge will enable us to extend our interpretation of the models which underlie the knowledge-base as they are exposed to processing a variety of real speech data.

Results of encoding some primary speech knowledge into the system were reported in Ran and Millar (1991). In that paper we showed that pre-classification of speech into vocalic and non-vocalic, and into dynamic and pseudo-static portions prior to attempting a data-driven pattern classification, enabled enhanced classification performance. The present study extends that work using the same philosophy of breaking down the speech according to its known sub-structure and creating simple sub-recognisers which operate in parallel. The outputs of these sub-recognisers may subsequently be combined. Investigations of this post-processing are currently been carried out.

In this paper we create modules suitable for inclusion as sub-recognisers in a phoneme recognition system. Each module is designed to recognise one phonetic feature. The definition of the set of features are mainly based on Jakobson et al (1961). Jakobson et al defined "distinctive features" on the bases of speech production and acoustic spectral energy distribution. The features served as their unique description of the speech sounds with the aim of providing an universal description of speech sounds across languages. An example of the use of such features in determining the phonetic nature of a segment of speech was illustrated by Blumstein and Stevens (1979) who reported that the invariant cues for place of articulation could be extracted from the beginning 26ms of a stop consonant release. Their findings that these cues are represented by the gross shape of the spectrum correlate with some of Jakobson et al's features.

The aim of the current feature extraction experiments was to determine the feasibility of extraction of some features described theoretically by Jakobson et al with some extensions, to see how well this extraction could be applied across speakers, in order to apply it later as the first level of a modular system of phoneme recognition .

SPEECH DATA CORPUS AND ANALYSIS

We used *[stop][vwl][d]* sounds as the speech material for our experiments. Where *[stop]* represents the stop consonants, and *[vwl]* represents the eleven Australian nominally monophthongal vowels that is: *[stop]*=[p t k b d g] and *[vwl]*=[i ɪ ɛ æ ɑ ʌ ɒ ɔ ɷ u ɜ]. We have four speakers and five repetitions from each speaker speaking these *[stop][vwl][d]* sounds.

The data was hand segmented and labeled. Each word was segmented into a Voice Bar interval (if any), a burst onset interval, a transition interval from the stop consonant to the vowel, a pseudo-static vowel interval, and a transition interval from vowel to the final [d].

The criterion of segmentation based on visual inspection of the spectrograms of CVds is described below:

- (For *[Voiced stop][vowel][d]* words) The starting point of the prevoicing; The indication of this point is the presence of the voicing bar;
- End point of the prevoicing and starting point of the burst; the indication of this point is the ending of the voicing bar or/and the starting of the broad-band noise;
- End point of the burst and starting point of the transition from the stop to vowel; the indication of this point is the ending of the broad-band noise of high energy or the appearance of a relative clear structure of formants;
- End point of the transitional portion and starting point of the pseudo-static portion of the vowel; the indication of this point is the starting point of the relatively static formant structure;
- End point of the pseudo-static portion of the vowel and the starting point of the transitional portion from vowel to *[d]*; the indication of this point is the ending of the static formant structure and the starting point of the dynamic formant structure;
- Starting point of the *[d]* closure; the indication of this point is the absence of any formant energy above and presence of the voicing bar for the *[d]*.

The speech material used for this paper are the prevoicing interval, the burst onset interval, and the pseudo-static portion of the vowel. The signal was passed through a Hamming window of frame-length 12.8ms, and then 13 Linear Predictive Cepstral Coefficients (LPCC) for each analysis frame were calculated. This was repeated with 50% overlap between frames.

EXPERIMENT DESIGN

The definition of the set of phonetic features was mainly based on Jakobson et al (1961), with some extensions. We defined the set of features having the task of phoneme classification of the speech material (*[stop][vwl][d]*) in mind, that is we included only the features which are useful for this classification task. The set of features are 'Acute', 'Compact', 'Diffuse', 'Flat', 'Grave', 'Lax', 'Non-Vocalic', 'Plain',

Feature	*Description*
Acute	Upper side of the spectrum predominates
Compact	Relative predominance of one centrally located formant region (or formant)
Diffuse	One or more non-central formants or formant regions predominate
Flat	A set of formants or even of all the formants in the spectrum shifts downward
Grave	Lower side of spectrum predominates
Lax	Shorter sound interval and lower energy comparing with Tense
Non-Vocalic	Having more than one periodic source whose onset is abrupt
Plain	No shift of formants
Tense	Longer sound interval and a larger energy comparing with Lax
Vocalic	Having a single periodic source whose onset is not abrupt
Voice Bar	Presence of low frequency spectral energy
Voiced	Superposition of a harmonic sound source upon the noise source
Voiceless	Having noise source only

Table 1: Brief Description of the Features

'Tense', 'Vocalic', 'Voice Bar', 'Voiced', and 'Voiceless'. Each of the features is described briefly in Table 1.

For the design of the experiments of the extraction of the phonetic features , we take into account the following points:

- Independency: At this level of the recognition system, we need the extraction of every individual feature to be as independent as possible.
- Performance: We need a design which gives us accuracy of extraction for the test speaker to be as high as possible.

As described in the introduction section, the system reported in this paper can be seen as the first level of a phoneme recognition system. This part of the system comprises modules which are sub-recognisers where each module is designed to recognise a single phonetic feature. Each module was implemented as a Multi-Layer Perceptron (MLP)(Rumelhart, Hinton and Williams 1986). Each MLP consists of several layers of nodes : an input layer of 13 nodes, one or more hidden layers of variable number of nodes, and one output layer of two or three nodes. Output layers with two nodes are used in MLPs which encode evidence for the feature being 'on', and 'off', and those with three also encode evidence for the feature being 'irrelevant'. Fully-connected MLPs were used in which each node in one layer is connected to all the nodes of the adjacent layer. The selection of the number of hidden layers and the number of nodes in these layers defines the architecture of the MLP.

Each module was first of all 'trained' in order to encode the evidence for its feature within its inter-node weights, and then it was 'tested' by applying it in recognition mode to independent data. One speaker was chosen at random to be the 'test' speaker, and the other three were used to provide 'training' data.

In the training phase the 13 LPCCs of the selected speech intervals were presented to the input layer of the MLP of each module frame by frame, while the output was clamped to the corresponding target value of each of the features. The back-propagation training algorithm was used to compute the weights by repeating this process until a minimum in the error surface was found which could not be reduced further. Tables 2 and 3 provide a summary of the target feature values for each of the phonemes, where '+' means the feature is "on" and '-' means the feature is "off", '*' means the feature is "irrelevant" (this only serves for balancing the amount of training data in each class). These data were largely derived from Jakobson et al (1961), Mitchell (1962), and Hyman (1975).

Features	*Vowels*										
	i	ɪ	ɛ	æ	ɑ	ʌ	ɜ	u	ɷ	ɔ	ɒ
Vocalic/Non-Vocalic	+	+	+	+	+	+	+	+	+	+	+
Compact/Diffuse	-	-	-/-	-/-	+	+	-/-	-	-/-	-/-	+
Grave/Acute	-	-	-	-	-	-	-/-	+	+	+	+
Flat/Plain	-	-	-	-	-	-	-	+	+	+	-
Tense/Lax	+	-	-	-	+	-	+	+	-	+	-
Voiced/Voiceless	+	+	+	+	+	+	+	+	+	+	+
Voice Bar	-	-	-	-	-	-	-	-	-	-	-

Table 2: Feature Values for Vowels

Features	*Stop Consonants*						
	p	b	t	d	k	g	Voice Bar
Vocalic/Non-Vocalic	-	-	-	-	-	-	*
Compact/Diffuse	-	-	-	-	+	+	*
Grave/Acute	+	+	-	-	+	+	*
Flat/Plain	-	-	-	-	-	-	*
Tense/Lax	+	-	+	-	+	-	*
Voiced/Voiceless	-	+	-	+	-	+	*
Voice Bar	-	-	-	-	-	-	+

Table 3: Feature Values for Stop Consonants

In order to define the optimal architecture of MLP for each feature, we trained a group of MLPs with different architectures of one or two hidden layers of different numbers of hidden units. For each candidate architecture, we trained with 100 initial conditions, in order to allow the training to start from a different position for finding the minimum. This strategy helps us to overcome the possibility of using an initial condition which leads to a local minimum.

In the test phase, the same type of the speech material from the test speaker was used. The 13 LPCC of the test material was presented to the trained MLPs of each module, and their output was compared with the expected target feature values allowing a recognition score for each MLP to be obtained. The best architecture for each feature was determined by choosing the architecture which provided the best recognition score for that feature in the testing phase.

RESULTS

The best architecture for each of the features is given in the Table 4, where the notation of the architecture is <number of the input nodes>-<number of nodes in first hidden layer>[-<Number of nodes in second hidden layer>]-<Number of the output nodes>. For example, 13-14-3 means there are 13 input nodes, 14 nodes in the hidden layer, and 3 output nodes.

Because of the different nature of the vowels and the bursts, we report the recognition score of the features in vowels and bursts separately. Note however that the modules were trained with data from both vowel and burst onset intervals, except for the modules representing Voiced and Voiceless features which were trained with data from burst intervals only . Table 5 gives the result for testing by presenting vowel materials only, and Table 6 gives the result for testing by presenting data from burst intervals only. Tables 5 and 6 represent the results obtained from the best architecture for each of the features.

DISCUSSION

- Analysing the results in the Table 5 and Table 6, we can see that for some features the recognition score is not very high. The reason be because of the different nature of the vowels and the burst, their acoustic characteristics for the features are different. This suggests training the feature modules for vowels and stops separately. Preliminary results of such separate extraction shows a 10% improvement on average.

- The recognition result of the features Lax and Tense are relatively low. This most likely is due to the fact that these features relate to the duration of the phoneme, and in a frame by frame analysis the duration is not captured. A model taking the duration into account should be investigated.
- The result reported in this paper was based on one randomly selected test speaker. A complete rotation of the speakers for selection of the test speaker in order to avoid speaker-dependent results is in progress.
- There has been a continuing debate about the existence of invariant phonemic cues in the speech signal. Blumstein and Stevens (1979) insist that such cues exist, and demonstrated this by extracting cues for place of articulation from the beginning 26ms of stop consonants in CV and VC context. They also showed that these cues are encoded in the gross shape of the spectrum sampled at the consonantal release. Their features were 'Diffuse-Rising', 'Diffuse-Falling' and 'Compact' which correlate with Jakobson et al's 'Acute', 'Grave', 'Compact', and 'Diffuse' ('Diffuse-Rising' to 'Diffuse' and 'Acute'; Diffuse-Falling' to 'Diffuse' and 'Grave'; 'Compact' to 'Compact'). They reported an average rate of extraction of these features of 85%. Our results, based on a relatively small amount of data give similar support to the invariant presence of these features.

REFERENCES

Blumstein, S. E. and Stevens, K. N. (1979). *Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants.* J. Acoust. Soc. Am. 66(4), Oct. 1979, pp1001-1017.

Huckvale, M. (1990). *Exploiting Speech Knowledge in Neural Nets for Recognition*, Speech Communication, vol. 9, pp1-13.

Hyman. L. M. (1975). *Phonology, Theory and Analysis*, (Holt, Rinehart and Winston), 1975.

Jakobson, R. Fant, C. G. M. and Halle, M. (1961). *Preliminaries to Speech Analysis, The Distinctive Features and their Correlates*, Technical Report No.13 of the M.I.T. Acoustics Laboratory, (The M.I.T. Press).

Mitchell, A. G. (1962). *Spoken English* (New York.St. Martin's Press), 1962.

Ran, S. and Millar, J. B. (1991). *Phoneme Classification using Neural Networks Based on Acoustic-Phonetic Structure*, 2nd European Conference on Speech Communication and Technology, Genova, Italy, pp125-132.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, (MIT Press, Cambridge, MA).

Zue, V. W. (1985). *The Use of Speech Knowledge in Automatic Speech Recognition*, Proceedings of the IEEE, vol. 73, No. 11, Nov. 1985.

Feature	*Architecture*
Acute	13-14-3
Compact	13-12-3
Diffuse	13-12-3
Flat	13-12-3
Grave	13-10-3
Lax	13-12-3
Non-Vocalic	13-6-4-3
Plain	13-8-4-3
Tense	13-12-3
Vocalic	13-6-4-3
Voice Bar	13-8-3
Voiced	13-8-2
Voiceless	13-8-2

Table 4: Architecture of the modules

Feature	*Rate*
Acute	83.8
Compact	90.1
Diffuse	73.8
Flat	89.6
Grave	84.1
Lax	80.8
Non-Vocalic	98.8
Plain	88.5
Tense	79.4
Vocalic	94.9
Voice Bar	100.0

Table 5: Recognition Rate of the Features for Static Vowels

Feature	*Rate*
Acute	82.1
Compact	90.4
Diffuse	79.3
Flat	97.8
Grave	77.2
Lax	54.1
Non-Vocalic	72.8
Plain	98.1
Tense	54.9
Vocalic	76.9
Voice Bar	98.3
Voiced	85.9
Voiceless	83.6

Table 6: Recognition Rate of the Features for Burst of the Stop consonants