PHONEME-LIKE MODEL OF SPEECH SIGNAL

S. Noreika and A. Rudžionis

Technological university, Kaunas, Lithuania

ABSTRACT

Phoneme-like representation of speech signal for single speaker isolated word recognition is discussed. Speech signal is divided into transitions and stationary parts by estimating a spectral instability function. The number of these parts is close to the number of phonemes. The comparison of reference and test patterns is based on processing of the similarity matrix.Well known dynamic time warping technique as well as our original technique are used. The recognition error rate is 0,9% (vocabulary size - 200 words, memory requirement - less than 40 bits per word).

INTRODUCTION

The problems of speaker independent speech recognition are well known. The employment of *a priori* information at the phonetic level is supposed to be the effective mean for speech recognition. New methods for phonemes detection [1] ensure high accuracy and error rate several times less as compared to other methods widely used in speech recognition technology. Nevertheless, the algorithm of the phonetic recognition of words is not clear enough. Our purpose is to discuss a phoneme-like model for single speaker speech and to evaluate the main parameters responsible for recognition results.

According to our model transitions and stationary parts of speech signal are detected. The number of these parts is close to the number of phonemes. It is achieved by estimating a spectral instability function. Extremal values of this function after filtering and thresholding procedures correspond to phonemes. On the stage of comparison of reference and test patterns we tried to find estimations, which were more efficient as compared to, e.g. dynamic time warping (DTW) technique estimation. To evaluate coarticulation phenomena, multiple reference patterns per phonetic unit were used.

Model testing resulted in 0.9% error rate of speaker-dependent recognition of 200 words. The phonetic transcription of a word was its reference and less than 40 bits of memory was required for its storage.

2. PHONEME-LIKE REPRESENTATION

2.1. Speech signal segmentation

The task is to divide speech signal into transitions and stationary parts. It is desirable that the number of these parts was close to the number of phonemes. In other words, a phoneme-like model is required. The metod of selecting transition and stationary frames is based on the estimation of a spectral instability function.

Let $\tilde{S}_{k,l}$ represents a set of logarithmic spectral vectors of speech signal, where k denotes the discrete time instant and $l=1,\dots,L$ denotes the number of a spectral component. The spectral instability function may be defined as follows:

$$\beta_{k,l}=\sum_{n=-n_o}^{n_o} n\,\tilde{S}_{k+n,l}\ , \qquad (1)$$

where $[-n_o,n_o]$ is the interval of spectral instability estimation, $n_o=2$.

The main segmentation function is

$$\beta_k=\frac{1}{L}\sum_{l=1}^{L}|\beta_{k,l}|\ . \qquad (2)$$

The maxima of this function are related to transitions while the minima are related to stationary parts of signal. To eliminate extremal values of the segmentation function which are related to local fluctuations of spectral parameters, filtering and thresholding procedures are adapted. The number of consecutive pairs "transition-stationary part" γ characterizes the extent of compression. Ideally, γ should be equal to the number of phonemes in a word.

2.2. Representation of feature vectors

Spectral instability coefficients and spectral parameters are used for description of transitions and of stationary parts accordingly. Both reference and test patterns may be represented as follows:

- feature vector consists of a set of successive frames, corresponding to transitions and stationary parts (the first phoneme-like model PLM1);
- transition and stationary part following are treated as a single component of a feature vector (the second phoneme-like model PLM2);
- vector quantization (VQ) may be applied to PLM1 or PML2.

2.3. Phoneme verification model

The modification of the phoneme like model is possible. We call it the phoneme verification model (PVM). The main distinguishing features of the PVM are: (1) the phonetic transcription of a word is its reference, (2) a database characterizes each element of the phonetic alphabet used, and (3) transitions are not used.

A database contains a set of spectral parameters of each phonetic element. Several versions may be used for representation of each phonema. The clustering technique is very suitable for this purpose.

3. Comparison of test and reference patterns

One of two matrixes can be used for the comparison of reference and test patterns: (1) a matrix of local distances and (2) a matrix of local similarities. We consider a similarity matrix is preferable to a distance one. A similarity matrix is supposed to have more information than DTW algorithm uses. Let d_{ij} is an element of a distance matrix, then an element of a similarity one is defined as:

$$\sigma_{ij}=\begin{cases} c_{ij}, & \text{if } c_{ij}>0 \\ 0, & \text{if } c_{ij}\le 0 \end{cases}, \quad (3)$$

where $c_{ij}= d_0 - d_{ij}$ and d_0 is some constant. Several measures of coincidence between reference and test patterns were investigated, they are presented in the following section.

4. EXPERIMENTAL RESULTS

4.1. Speech material

The speech material for testing PLM1 and PLM2 was recorded by one male speaker who uttered a 100-words vocabulary 10 times. Spectral analysis of the incoming signal was carried out by a bank of 8 analog bandpass filters. All the channels were sampled every 10 ms by a 8-bits analog/digital converter. The vocabulary consisted of 794 graphic simbols, i.e. on the average one word consisted of 7.94 letters. The extent of compression of various segmentation algorithms was evaluated on the base of this figure. In the recognition experiments the reference and test patterns were chosen according to the "leave-one-out" procedure, obtaining a total of 9000 tests. In some experiments only part of these tests was used.

4.2. PLM1 and PLM2 testing

Several variants of recognition were investigated. The first two methods were the usual DTW methods on the basis of a local distance matrix (V1) and of a local similarity matrix (V2). The third variant V3 differed from V2 by the normalization of the integral similarity measure according to the average duration of the reference and test patterns. The variants V4 and V5 are like the variants V2 and V3, but the formers use only three side-by-side diagonals of the similarity matrix having the largest similarity. The logical processing of elements belonging to these diagonals is the essence of the sixth variant V6. And finally, the seventh variant V7 is the modification of the variant V6, including the segmentation errors correction. Feature vectors for the variant V1 are represented according to PLM1, while the other variants use representation according to PLM2. The results are presented in Table 1, where N_t is the number of test patterns and ρ is the recognition error rate.

Our model ensures a high extent of compression and the number of detected phonemes γ is close to the average number of letters (7.94). Generally, the recognition error rate is inversely related to the extent of compression. The normalization of the integral similarity measure and the employment of diagonals reduce the recognition error rate. The variants V6 and V7 give the best results and these results are achieved without using DTW algorithm.

4.3. Vector quantization

A 128-element codebook was generated for PLM1 (memory requirement was about 100 bits per reference) and for PLM2 (memory requirement was about 50 bits per reference). The recognition results are shown in Table 2. Naturaly, VQ reduces the recognition accuracy, nevertheless, the rezults are high enough on condition that such an extent of compression is used.

4.4. PVM testing

To test this model, a 200-words vocabulary was used. As mentioned above, the phonetic transcription of a word was its reference. In the recognition experiment vocabulary was read 7 times, i.e. the total number of tests was 1400 words. The database was formed by clustering speech material containing 50-100 repetitions of each phonema. Some phonetic units were considered as one phonema, e.g. /p,t,k/ or /b,d,g/, so only 16 phonetic units were used. Hence it follows that memory requirement was only $4m$ bits per reference, where m is the number of phonemes in a word. The recognition was carried out according to variant V6, except that only two diagonals were used. The rezults are shown in Table 3. The model gives the promissing rezults. They are conditioned mainly by the use of *a priori* information about phonemes and by the proper processing of the similarity matrix. Note the main attractive features of this model: (1) practically extremal compression of speech is achieved, (2) once the database have been formed it may be used with any vocabulary, (3) the amount of similarity calculations does not depend on vocabulary size and (4) vocabulary can be changed easily.

CONCLUSION

The models used here ensure high extent of compression of speech signal without degradation of useful information. Recognition of 200 words showed that recognition error rate was 0,9% and memory requirement was less than 40 bits per reference. In the future these models are supposed to be used for speaker-independent speech recognition.

REFERENCE

[1]DOMATAS, A.and RUDŽIONIS, A. "Towards more reliable automatic recognition of the phonetic units", in this issue.

Table 1: Comparison among various variants of recognition

Variant	V1	V2	V3	V4	V5	V6	V7
N_t	9000	2700-3300				9000	
γ	9.1-7.4	7.2				7.2	
ρ,%	3.4-6.0	6.0	3.4	4.3	2.3	1.6	0.87

Table 2: PLM1 and PLM2 testing results with VQ

N_t	ρ,%		
	Without VQ	Memory requirement, bits	
		100	50
900	0.5	1.5	1.9

Table 3: PVM testing results

Number of clusters per phoneme	8	6	4
ρ,%	0.9	0.9	1.8