TOWARDS MORE RELIABLE AUTOMATIC RECOGNITION OF THE PHONETIC UNITS

A. Domatas and A. Rudžionis

Technological university, Kaunas, Lithuania

ABSTRACT

This paper is concerned with speaker-independent phoneme recognition in isolated words. We try to evaluate the influence of coarticulation and to create an economic and effective phoneme recognition method which estimates the correlations among features. An adequate evaluation of transitions between phonemes and application of a dichotomization-based (D) classifier permits to decrease the recognition error for several times as compared with widely used Euclidean (E) classifier.

1. INTRODUCTION

Speaker-independent recognition is so far related to great problems. Comparison of effectiveness of methods widely used [2] does not show essential difference among them. Moreover, the E classifier proves to be equal to other methods. Usually a phoneme is represented by features in its stationary part. Good results are presented in [1] with inclusion of dynamic features when discriminating between nasals. However, these results are obtained on reference set only. Here an approach of automatic estimation of coarticulation and the classifier using an a priori information effectively are proposed.

2 . PHONEME RECOGNITION

2.1. Use of coarticulation

Speech signal is represented by consistently following spectral vectors $S(k)=\{S_n(k)\}$, where $n=1,\dots,N$ is the number of spectral component, and k is the number of spectral vector.

The instability of every spectral sample is:

$$E_n(k) = \sum_{s=-2}^{2} s\, S_n(k+s) \qquad (1)$$

The instability of spectral vector k is:

$$e(k)= \frac{1}{N} \sum_{n=1}^{N} | E_n(k)| \qquad (2)$$

Spectral vector $S(m)$ where $m=\underset{k}{\operatorname{argmin}}\, e(k)$ corresponds to the stable part of a phoneme and vector of instabilities $E(h)$ where $h=\underset{k}{\operatorname{argmax}}\, e(k)$ corresponds to the transitionary part. The logical rules are applied to exclude false extrema. In our experiment, the vector of initial features X of a phoneme is formed in the three ways:

-from spectrum $X=\{S(m)\}$(SP);

-from spectral vector and vector of instabilities $X=\{S(m),E(h)\}$(SPI);

-from spectrum and consistently following vectors of instabilities $X=\{S(m),E(h-2),E(h-1),\dots,E(h+2)\}$ (SPCFI).

Let I, $I \geq N$ to be the number of initial features.

2.2. Dichotomization between phonemes

A linear function to discriminate between phonemes s and t is used:

$$g(s)=\sum_{i=1}^{J} W_i^{st} | x_i^{st} - \bar{x}_i^{s} | + Q^{st} \qquad (3)$$

where $x^{st} \in X$ represents a vector of selected features of unknown test pattern when distinguishing between s and t. Respectively $\bar{x}^{s}$ represents a vector of selected reference features of phoneme s, W is a vector of weights, J is the number of selected features, Q is a threshold.

For dichotomization of every pair of phonemes the own threshold and sets of features and weights are calculated. During the training we calculate averages $\bar{x}_i^{s}, \bar{x}_i^{t}$, $i=1,\dots,I$, and correlation matrixes C^{s}, C^{t}. The Gaussian distribution of features is supposed. Features are ordered according to the decrease of interphoneme distances

$$d_i^{st} = \frac{| \bar{x}_i^{s} - \bar{x}_i^{t} |}{\sigma_i^{s} + \sigma_i^{t}} \qquad (4)$$

where σ^2 is the variance. Weights W_i^{st} are calculated by using an iterative procedure to minimize the probability of misclassification p_j^{st} (j is the number of weights already defined). The procedure estimates correlations among features.

Number of selected features J is computed by:

$$J=\underset{1 \leq j \leq I}{\operatorname{argmin}}\, \bar{p}_j^{st} \qquad (5)$$

where $\bar{p}_j^{st}$ is expected probability of error. $\bar{p}_j^{st}$ depends on training set size, on j and on p_j^{st} and is defined from the tables in [3].

2.3. Dichotomization-based classifier

Output of an elementary dichotomie $O(s)$ is denoted by:

$$O(s)= \begin{cases} 1, & \text{when } g(s) \geq 0 \\ 0, & \text{when } g(s) < 0. \end{cases} \qquad (6)$$

Respectively, output $O(t)$ is

$$O(t)=1-O(s). \qquad (7)$$

Here two approaches are used to get the final result:

-consistent elimination(D1): class t is excluded from the list of classes considered if $O(t)=0$, and class s is compared to the next class from the list. The result is the class remained after $S-1$ comparisons where S is the number of classes.

-voting(D2): the result is class v defined by:

$$v=\underset{1 \leq s \leq S}{\operatorname{argmax}}\, O(s) \qquad (8)$$

3. EXPERIMENTS AND RESULTS

3.1. Experimental conditions

-filter bank N=24 or N=8 (averaging 3 neighbouring) nonlinearly spaced channels;

-interval of spectral frames 10 ms;

-sample quantization 8 bits.

3.2. Recognition of stationary vowels

The comparison of D, E and Mahalanobis (M) classifiers was performed to estimate their effectiveness. The speech material consisted of phonemes from words /a/,/o/, /u/,/i/ spoken by 12 males. (4800 patterns). The error rate of reference set recognition (C-examine) and "leave-one-out" recognition (L-examine) is shown in Table 1. Results show that D classifier reduces error rate for more than 4 times in comparison to E one and needs less training than M one: for 11 speakers D classifier led to similar error rate for both C and L examines.

In this experiment, D and E classifiers required the similar recognition time.

3.3. Recognition of coarticulated /m/,/n/,/v/,/l/

This experiment was performed to investigate the effect of inclusion of dynamic features. The diphones consonant-vowel were selected from words, where vowel was {/a/,/u/,/i/}. 11 male speakers took part in this experiment. Reference and test sets consisted both of 220 patterns of every coarticulated consonant (2640 patterns in all). The error rate of the test set recognition is shown in Table 2. Results presented suggest that correct selection of features and use of D classifier provides for recognition error rate less than 4% for all three cases:

-efficient discrimination of these 4 consonants in context with /a/ is achieved by using stationary part of consonant only;

-in context with /u/ it is necessary to add dynamic features in transition between consonant and vowel;

-even in the most complicated situation when discriminating among soft consonants, the use of several vectors in transition leads to very low error rate.

3.4. Additional superiority of D classifier

The influence of transmission channel on recognition error rate was examined. One male speaker pronounced 100 patterns of every nasal /m/, /m'/,/n/,/n'/ in combinations nasal-vowel during training. The hard nasals were pronounced in /a/ context, the soft nasals /m'/, /n'/ - in /i/ context. 100 patterns of every nasal were used for test set by using the same microphone , and by using microphone of another type. Feature system SP was used. The recognition error of test sets is shown in Table 3. Results show that D classifier is less sensitive when changing the properties of the transmission channel in comparison to E one.

3.5. Automatic labeling of isolated words

The aim of the experiment is the comparison of automatically obtained transcriptions of words to manually formed references. 20 phonemes (50 phonetical subclasses) were used. The alphabet consisted of Lithuanian phonemes except /r/. Stops /p/, /t/,/k/ and /b/,/d/,/g/ were united to "unvoiced stop" and "voiced stop" respectively. The labeling process includes two steps. First, the feature system SP is applied. Second, if connection sonant-vowel, nasal-vowel or mixed-vowel is fixed, system SPI is used for more accurate definition of a consonant according to the vowel recognized. 11 males took part in forming reference set, each subclass consisted of 200-1500 patterns. Test set was formed from 50 words spoken twice by 10 males. Average word length was 7.0 phonemes. The correct transcription of a word was fixed if it adequately coincided with the transcription of its reference.The test led to 32% correct transcriptions of words for D classifier and to 6.2% for E one correspondingly.

4. CONCLUSIONS

We have presented two methods to improve phoneme recognition. Inclusion of dynamic features into representation of phonemes provides for significant descrease of recognition error rate.

Dichotomization-based classifier offers the following advantages:

-inclusion of essential features only for dichotomization between phonemes;

-selection of feature set guarantying minimum probability of dichotomization error;

-immaterial influence of transmission channel because of effective application of correlations among features;

-less training set necessary to form representative references in comparison to Mahalanobis one;

-less recognition time required in comparison to Mahalanobis one;

-lesser error rate for several times in comparison to Euclidean one.

5. REFERENCES

[1] HARADA, T., KAWARADA,H., (1986),"High resolution frequency analysis of voices - feature extraction of /mu/ and /nu/" , *Bull. P. M. E. (I. I. T.) - No.* 58. *P.* 1 - 10.

[2] RABINER,L.R., SOONG,F.K. (1985), "Single-frame vowel recognition using vector-quantization with several distance measures", *AT and T Bell Lab.Techn.Journ.* -64. *-No.* 10. *-P.* 2319-2330.

[3] RAUDYS, Š., PIKELIS, V. (1975), "Tabulating of the probability of misclassification for the linear discriminant function",*Vilnius, Statistical methods of control*. *-No.* 11. - *P.* 81-119.

Table 1

Percentage error rates for vowels /a/,/o/,/u/,/i/

Classifier	Number of features	Examine	NS=1	NS=4	NS=11(12)
E	I=8	C L	5.3 15.6	8.2 11.5	12.2 13.2
M	I=8	C L	0.6 14.3	0.7 10.6	1.8 6.6
D1	$J \leq 4$ (I=8)	C L	- -	1.0 3.9	2.6 2.8

NS is the number of speakers used for reference forming

Table 2

Percentage error rates for coarticulated /m/,/n/,/v/,/l/

Classifier	Method of phoneme representation	Number of features	Vowel of diphone /a/	/u/	/i/
E	SP	I=24	11.0	16.1	18.8
E	SPI	I=48	5.0	9.5	12.7
D2	SP	$J \leq 12$ (I=24)	2.9	8.3	10.0
D2	SPI	$J \leq 12$ (I=48)	1.3	3.8	6.7
D2	SPCFI	$J \leq 12$(I=144)	-	-	1.7

Table 3

Testing of microphone change (percentage error rates)

Classifier	The former microphone	Another type microphone
E	4.0	13.9
D1	1.5	2.6