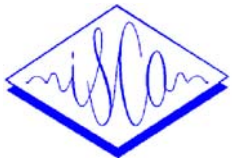

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PHONEME RECOGNITION IN FIXED CONTEXT USING REGULARIZED DISCRIMINANT ANALYSIS

A. Rudžionis, V. Rudžionis

Kaunas University of Technology, LT-3006 Kaunas, Lithuania
Vilnius University, LT-2734 Vilnius, Lithuania
Alrud@mmlab.ktu.lt

10.21437/Eurospeech.1999-604

ABSTRACT

Speaker independent discrimination of four confusable consonants in the strictly fixed context of six vowels is considered. The consonants are depicted by features of consonant's stationary part and changing rate of features (delta features) in transition from consonant to the following vowel.

The mel frequency cepstrum (MFCC), linear prediction cepstrum (LPCC), recursive filter (F12) features and set of discriminants were evaluated seeking for better phoneme discrimination. It is postulated that Gaussian mixture capabilities are similar to k-means (kMN) capabilities and several discriminants including regularized discriminant analysis (RDA) were analyzed too.

The experiments showed that the discrimination error averaged per environments of six vowels decreases from 23.3% using kMN to 7.0% using RDA for the best F12 features. Consonant discrimination error rate decreases from 21.6% to 3.6% in the open vowel context and from 27.9% to 11.4% in closed vowel context.

1. INTRODUCTION

The continuous density hidden Markov model (CD HMM), neural networks (NN), N-gram language models were among the mostly important factors that marked progress in the speech recognition during last decade. Despite this lots of problems still needs to find better solution (understanding of the basic phonetical units, robustness to channel variability, background noise, etc.) [1]. It is expected that improved phoneme discrimination should lead to more reliable overall performance of speech recognition systems.

There were many attempts to increase phoneme discrimination accuracy. The context-dependent phoneme models, subspace principle applied to 26 filterbank log energies and Bhattacharyia distance allowed to achieve 96.6% nasal discrimination rate for high quality ISOLET data [2]. The E-set recognition performance droped from 95% to 71.8% when telephone quality OGI database were tested.

The mixture of cepstrum trend functions has been used for TIMIT database phonetic classification [3]. The minimum classification error learning criteria led to 83.48% correct phoneme recognition.

The linear discriminant analysis (LDA) places most emphasis to recognition borders, mainly ignoring easily recognized tokens during learning. The whole-word adaptive LDA achieved 4% error rate on BTL E-set [4].

The objectives of this study was to search for the ways to improve discrimination of acoustically similar phonetic units trying to compare performance of different discrimination methods on same task. Aiming this there was performed analysis of the consonant discrimination in the strictly defined phonetic environment, it means. consonants preceding a given vowel were considered. The consonants are depicted by features of their stationary parts and by delta features in transitions from consonants to the following vowel. The stationary part and transitional region are determined automatically by segmentation algorithm [5] and are only approved manually. In this way each consonant is represented by fixed length vector and it is expected that this should allow to evaluate potential capabilities of local discrimination. The question how to implement the discriminant to speech recognition algorithm such as HMM has not been considered.

The regularized discriminant analysis (RDA) combined with perceptron [6] was used for phoneme classification purposes. The RDA is a modification of LDA where sample estimates of classifier parameters are regularized through their eigenvalues and eigenvectors.

We postulate that local discrimination capabilities of the most popular continuous density HMM with Gaussian mixtures model should be similar to the k-means (kMN) classifier capabilities. The Gaussian mixtures uses sample estimates of features covariance but as has been seen in various studies there were observed no significant performance improvement comparing recognition accuracy with full covariance matrix and sufficient number of mixtures with diagonal covariance matrix. So here is paradigm to analyze kMN and RDA.

Our previous experience using recursive filter analyzer showed some advantages comparing with cepstrum and other features [7] (including FFT based bandpass filters) when discriminating the confusable phonetic units. Here are compared mel frequency

cepstrum (MFCC), linear prediction cepstrum (LPCC) and recursive filter features (F12).

Further paper is composed as follows: in section 2 we present feature extraction and discrimination methods, section 3 describes database used in this study, section 4 presents results of experiments while in section 5 some conclusions and comments are provided.

2. FEATURES AND DISCRIMINANTS

Features

The speech data were recorded at 10 kHz sampling rate using 8 bit per sample resolution with low cost electret microphone.

Speech analysis was performed by calculating recursive filters and cepstrum parameters every 6.4 msec. Digital filterbank (F12) parameters consists from log energies of 12 Butterworth digital bandpass filters spaced in mel-scale and covering frequency region 0-4000 Hz. The 25.6 msec Hamming window has been applied to speech frame before cepstrum analysis. Linear prediction cepstrum (LPCC) coefficients were calculated from 12-th order LPC analysis parameters. Mel frequency cepstrum (MFCC) coefficients were derived by sine liftering of discrete cosine transform coefficients. DCT was applied to the outputs of 20 mel scaled triangular filters which were used to filter magnitudes of 1024 points FFT (zero padded 256 signal samples).

The simple and delta features were derived from recursive filters or cepstrum parameters. The simple features were obtained by averaging 7 - 11 consecutive frames of the filter or cepstrum parameters. The delta features in the meantime were obtained by evaluation of change rate of analysis parameters in the same time interval. The sum of the absolute values of the recursive filters (F12) delta features after some linear and nonlinear filtering was used as a segmentation function.

Phoneme discrimination features were obtained from simple and delta features. The minimums and maximums of the segmentation function were considered as labels of stationary and transitional regions respectively. The consonant is depicted as a single vector joining the simple features vector Cs on the consonant stationary region and the delta features vector Td on the transitional region from consonant to the following vowel.

The fig. 1 shows how Cs and Td were determined. The speech wave of the triphone vowel V1 - consonant C - vowel V2 together with sonogram, markers of the stationary - transitional regions and segmentation function are presented from top to bottom. The stationary region of consonant Cs is approved manually as a minimum of the segmentation function (lower column in the third row) analyzing this picture and listening a sound. The transitional region Td then is automatically marked at the following maximum (higher columns in the third row) of the segmentation function. This example was selected to show the difficulties of manual V1-C-V2 segmentation using only speech wave.

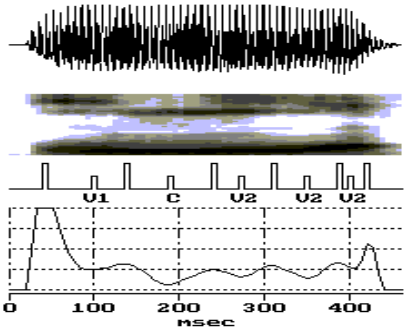

Fig. 1. Triphone vowel V1 - consonant C -vowel V2 segmentation. From top: a) speech wave, b) sonogram, c) stationary (lower columns) and transitional (higher columns) regions, d) segmenting function.

Joining the Cs and Td parts representing features tends to evaluate coarticulation. The consonant is represented by double size vector (e.g., 24 coefficients in single vector: 12 for stationary and 12 for transitional parts).

Discriminants

Four discrimination algorithms were used in this study: a) k-means (kMN), b) Fisher, c) dichotomy (DCH), d) several modifications of regularized linear discriminant analysis (RDA).

The kMN discriminant is based on Euclidean distance and clusterization of the training data.

As was mentioned above we postulated that local discrimination capabilities of Gaussian mixtures should be similar to the k-means (kMN) classifier performance. So there is paradigm to analyze kMN and other discriminants.

The Fisher discriminant (FSH) is well known tool for vector discrimination and is based on Gaussian pdf assumption. The dichotomy discriminant (DCH) includes subspace principle when the discriminant is optimized going from the best feature to more efficient using empirical adjusting.

A scaled matrix rotation - a form of regularized linear discriminant analysis (RDA) - has been applied for consonant discrimination purposes also. In the standard RDA only the eigenvalues are regularized while in scaled rotation approach the eigenvectors are regularized too. The singular value decomposition $\mathbf{S}=\mathbf{TDT}'$ describes sample covariance matrix $\mathbf{S}$ as a function of the eigenvalues and eigenvectors. The estimate of sample covariance matrix is transformed to regularized estimate by scaling eigenvalues and rotating eigenvectors in scaled rotation regularization method in scaled rotation method:

$$\mathbf{S}' = \mathbf{T}^{\alpha}(\mathbf{D}+\lambda\mathbf{I})^{-1/2}\mathbf{T}^{\alpha\prime}\mathbf{S}$$

The scaling (regularization of the eigenvalues) of the sample covariance matrix is controlled by the changes of the parameter λ and the rotation

(regularization of the eigenvectors) of the matrix is controlled by the parameter α.

The single layer perceptron has been used for classification also.

The different discrimination methods were used:

SRDA - standard regularization of the eigenvalues (only scaling parameter λ is optimized);

SLP - single layer perceptron is used to classify consonant feature vectors**;**

SR - scaled matrix rotation (both parameters λ and α are optimized);

SR + SLP - rotated feature vectors classified with single layer perceptron.

3. DATABASE

The speech corpora was collected with the aim to evaluate discrimination peculiarities of nasal and sonorant consonants in different phonetical environments.

The database consists of utterances of 4 consonants **(M, N, V, L**) which precedes 6 Lithuanian vowels (**A, O, E, Y, U, I**). The English words *mark, moon* and *mean* are similar examples when nasal **M** precedes vowels **A, U, I** respectively going from the open vowel **A** to the closed vowel **I**.

The each of 20 male speakers pronounced every consonant in the triphones vowel-consonant-vowel (V1-C-V2) 10 times where the second vowel was always stressed. The first vowel V1 and the second vowel V2 were the same.

Then simple Cs and delta Td features of consonant were determined in stationary and transitional regions with the procedure explained above (Fig. 1).

Phoneme discrimination experiments were carried out using leaving-one-out speaker (tested speaker rotation) technique to simulate speaker independence. It means that speech data of one speaker has not been used during training while testing has been performed on this data. Averaged recognition accuracy for all speakers is presented in the tables.

3. EXPERIMENTAL RESULTS

Below there are presented averaged per all speakers simple and delta features derived from recursive filter F12 parameters of two nasals **M, N** in the contexts of 3 vowels **A, U, I** (Fig 2). The upper row shows simple (Cs) while lower row delta (Td) features. Consonant features before vowels **A, U, I** are plotted from left to right.

It is obvious that nasal discrimination complexity grows going from the context of open vowel **A** to closed vowel **I**. There are no almost any remarkable differences between averaged features of two nasals on stationary consonant part in the latter case (right picture on the upper row). It is expected too that discriminants which take into account the correlation among features (all discriminants except k-means) could be more effective for open vowel context (left picture on the upper row). Features of the transitional part that are shown in the lower row has many properties of the following consonant vowel.

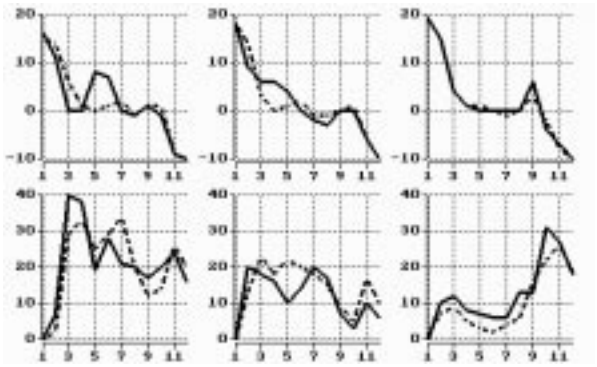

Fig.2 Stationary Cs (top) and transitional Td (bottom) parts averaged features for nasals **M** (solid line) and **N** (dashed line). From left to right - **A, U, I** contexts respectively.

We compared performance of k-means (kMN) and dichotomy (DCH) discriminants using recursive filter F12 features. Results are presented in the Table 1. kMN discriminant used 16 clusters in these experiments. The consonant discrimination error rate is presented for open vowel **A**, closed vowel **I** and averaged per contexts of 6 vowels.

Table 1. Consonant discrimination error rate (%) using F12 features and comparing k-means (kMN) and dichotomy (DCH) discriminants. The results are presented for open (**A**), close (**I**) and averaged per 6 vowels context.

Discriminant	Vowel after consonant		
	A	**I**	avr.
KMN	21.6	27.9	23.3
DCH	11.4	19.8	16.3

The averaged consonant discrimination error rate per all vowel environments decreased from 23.3% for k-means discriminant to 16.3% for dichotomy discriminant. This could be explained by the fact that dichotomy discriminant is able to use correlation between features while k-means discriminant don't. This decrease in error is at least not lower than difference between full covariance and multiple diagonals observed in many studies when recognition error rate is comparable. So we think it is reasonable to search for better suited than Gaussian mixture phoneme discrimination methods.

The second observation from Table 1 is as follows. Consonant discrimination error before the open vowel **A** is significantly lower than before closed vowel **I** especially when dichotomy discrimination method is used.

Taking into account earlier observed discrimination results we evaluated regularized discriminant analysis and scaled rotation maximization algorithms in the next experiment. The mel cepstrum (MFCC), linear prediction cepstrum (LPCC) and recursive filter (F12) features were evaluated also. Table 2 shows discrimination error rate of 4 consonants for open (**A**) and closed (**I**) vowel contexts when Fisher (FSH)

discriminant and several modifications of regularized discriminant analysis (RDA) were used.

Table 2. Consonant discrimination error rate using Fisher (FSH) and different regularized discriminant analysis methods (SRDA - standard regularized discriminant analysis, SLP - single layer perceptron, SRM - scaled rotation maximization, SR+SLP -rotated data and SLP)

Discriminant	Vowel **A** context			Vowel **I** context		
	LPC	MFC	F12	LPC	MFC	F12
FSH		-	14,7		-	22,2
SRDA	24,6	19,5	9,2	20,6	25,1	18,5
SLP	17,2	13,1	6,6	14,1	17,6	12,1
SRM	19,0	14,2	5,1	16,6	21,7	14,9
SR+SLP	15,2	8,7	3,6	11,9	16,0	11,4

It should be noted that recursive filter features F12 outperformed both mel cepstrum MFCC and linear prediction cepstrum LPCC. The experiment shows that Fisher discriminant (FSH) provides slightly bigger error than dichotomy discriminant (DCH) in the Table 1. DCH classifier incorporates subspace approach and only 2/3 of best features from 24 total number of features in vector were used.

But most substantial error reduction was achieved when regularized discriminant was combined with single layer perceptron (SR+SLP in lowest row Table 2). The discrimination error rates of 3.6 % in vowel **A** context and 11.4% in vowel **I** context were observed for F12 features.

Table 3 presents discrimination error rate for F12 features in the context of all 6 vowels.

Table 3. Consonant discrimination error rate using different methods of regularized discriminant analysis for 6 vowel contexts (F12 features).

	A	**O**	**E**	**Y**	**U**	**I**	avr.
FSH	14,7	7,1	18,0	20,3	22,1	22,2	17,4
SRDA	9,2	3,1	10,2	11,2	15,4	18,5	11,3
SLP	6,6	2,5	10,1	11,0	11,6	12,1	9,0
SRM	5,1	1,2	7,9	9,1	9,6	14,9	8,0
SR+SLP	3,6	1,1	9,1	8,5	8,5	11,4	7,0

As has been shown in Table 3 discrimination error decreases in a similar rate in all vowel contexts using regularized discriminant analysis together with single layer perceptron for classification. The averaged per all vowel contexts discrimination error was 7%.

5 . CONCLUSION

The confusable consonants preceding a given vowel were depicted by feature vector on consonant joined with delta feature vector on the transition to the following vowel. Evaluation of different types of features showed that recursive filter derived features outperformed cepstrum.

The regularized linear discriminant with single layer perceptron allowed to decrease average consonant discrimination error to 7% comparing with 23.3% for k-means discriminant. This decrease was especially remarkable in contexts of open vowels (from 21.6% to 3.6%) and less effective in closed vowel context (from 27.9% to 11.4%).

It is worth to note that very low (less than 2%) discrimination error was achieved earlier even for close vowel context when 24 analog bandpass filter features in 8000 Hz frequency region were used [7].

It is important to implement better phoneme discrimination capabilities to speech recognition algorithms. The precisely defined phoneme databases are desirable.

ACKNOWLEDGMENT

This study was partly supported by Lithuanian science foundation grants. We thank prof. S. Raudys for introducing to regularized discrimination analysis.

REFERENCES

[1] R.P.Lippmann. “Recognition by Humans and Machines: Miles to Go Before We Sleep”, Speech Communication, vol. 18, April 1996

[2] P.Loizou, A. Spanias. "High Performance Alphabet Recognition," IEEE Trans. on Speech and Audio Processing , vol.4, no 6, November 1996, pp. 430-445.

[3] R. Chengalvarayan, L. Deng. “Speech Trajectory Discrimination Using the Minimum Classification error Learning”. IEEE Trans. on Speech and Audio Processing, vol. 6, no 6, November 1998

[4] C.M.Ayer, M. Hunt, D.Brookes. “A Discriminatively Derived Linear Transform for Improved Speech Recognition.”, Proc. EUROSPEECH, Berlin, September 1993

[5] A.Rudzionis, ‘Recognition by averaged templates’, COST 249 "Continuous Speech Recognition Over The Telephone", Draft Minutes of the 1st Management Committee Meeting, Brussels, Belgium, 1994, pp. 41-47.

[6] V. Rudzionis, “Speech Recognition by Phonetic Units“, Ph.D. Thesis, Kaunas University of Technology, 1998.

[7] A.Domatas, A.Rudzionis. "Towards more reliable automatic recognition of the phonetic units," Proc. of the XIIth ICPhS, Aix-En-Provence, France, 1991, vol.4, pp 478-481.