Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification

Nick J.-C. Wang[1,2], *Wei-Ho Tsai*[1,3], *and Lin-Shan Lee*[2]

[1] Philips Research East Asia-Taipei, Taipei
[2] Graduate Institute of Communication Engineering, National Taiwan University, Taipei
[3] Department of Communication Engineering, National Chiao Tung University, Hsinchu
Taiwan, Republic of China

nick.wang@philips.com

Abstract

Eigen-MLLR coefficients are proposed as new feature parameters for speaker-identification in this paper. By performing principle component analysis on MLLR parameters among training speakers, the *eigen-MLLR coefficients* (EMCs) are derived as the coefficients for the eigenvectors. The discriminating function of the new EMC features based on the Fisher criterion is found to be ten times larger than that of mel-frequency cepstral coefficient (MFCC) features, for distinguishing speakers. The speaker-identification accuracy using the EMC features are shown to be significantly better than that using MFCC features, especially when the quantity of enrollment data is limited. It is also shown that properly combining MFCC and EMC features can achieve a significant error rate reduction on the order of 50%-60% as compared to using MFCC features alone.

1. Introduction

Speech recognition and speaker recognition are two complementary problems in speech signal processing. In speech recognition, the objective is to decode the linguistic message or phonemic information carried by the speech signals, and hence the variations caused by different speakers are considered harmful and should be removed or avoided as much as possible. Contrarily, speaker recognition aims to distinguish one speaker from the others, regardless of what they said, and hence the differences of voice characteristics among speakers must be exploited, while the phonemic differences for the speech signals are better ignored. Intuitively, speech recognition and speaker recognition should better use different signal traits as inputs. However, the most widespread feature parameters used to date in these two tasks are usually very similar, primarily based on spectrum-derived parameters, in particular the mel-frequency cepstral coefficients (MFCCs) [1]. As a result, in order to obtain the desired information – either speech content or the speaker identity, both tasks rely heavily on statistical models trained using large quantities of data. Such systems are therefore vulnerable and may be failure when the available speech data is limited. It is certainly desirable to extract only those acoustic properties for phonemic differences for the task of speech recognition, and extract only those acoustic properties for speaker variation for speaker recognition, if at all possible.

In speech recognition, a prevalent way to deal with the problem of speaker variation is speaker adaptation. It aims to reduce the acoustic mismatch between the training conditions and testing conditions by adapting the speaker-independent phonetic-acoustic models to a new speaker based on not too much data produced by the new speaker. Maximum Likelihood Linear Regression (MLLR) and eigenvoice are two important approaches in speaker adaptation. Both of them provide some approaches to handle the speaker variations, especially when only very limited quantities of enrollment data are available. In this paper, the aim is to explore new feature parameters carrying better information on speaker variation for speaker recognition. The well adopted Gaussian Mixture Modeling (GMM) technique [1] for speaker-identification tasks were used in the following experiments. Eigenvoice approach has been used in speaker recognition before [2], but in fast training of speaker-specific GMMs from a speaker-independent GMM via eigenvoice adaptation, which is quite different from what is presented below in this paper.

2. Eigen-MLLR coefficients

2.1. Maximum Likelihood Linear Regression

MLLR adaptation approach [3] is a well-accepted approach for speaker adaptation, which reduces the mismatch between the speaker-independent phonetic-acoustic models and the speaker-specific characteristics of the individual speakers by performing one or several affined matrix transformations on all mean vectors of the phonetic-acoustic models to approximate the speaker-specific characteristics. Apparently the coefficients in these MLLR matrices do carry the information of the speaker-specific characteristics, which might be used as features in speaker recognition. However, the number of free parameters in MLLR affined transformation matrices may be too large to be used for speaker recognition with limited quantities of training data. This problem may be handled by reducing the degree of freedom in MLLR parameters based on eigenvoice analysis as discussed below.

Paperid: 1353 Topic: 8 - Subm.: 2001-03-30 13:09:00 - Dist.: 2001-04-01 12:26:52 - Offset: 0 0

2.2. Eigenvoice analysis on MLLR matrix coefficients (eigen-MLLR)

The eigenvoice analysis [4] can be applied on the MLLR parameters [5][6], so as to preserve the advantage of efficient speaker characteristics of MLLR, but utilize the dimension reduction functions and variation maximization nature of eigenvoice. This concept is summarized below and referred to as *eigen-MLLR* here in this paper.

To begin with, the MLLR matrix coefficients for a specific speaker i are augmented and aligned as a long *'supervector'* w_i. This vector w_i is considered as a sample of a random vector $\mathbf{w}$ of speaker characteristics. When enough number of such samples are collected for a large number of speakers, say $i = 1, 2, ...K$, principle component analysis (PCA) [7] can be performed on the covariance matrix $C_{\mathbf{w}}$ of $\mathbf{w}$,

$$C_{\mathbf{w}} \quad = \quad E_{\mathbf{w}} \Lambda_{\mathbf{w}} E_{\mathbf{w}}^{\mathrm{T}}, \tag{1}$$

where $\Lambda_{\mathbf{w}}$ and $E_{\mathbf{w}}$ are the eigenvalue and eigenspace matrix, respectively.

The eigenvectors obtained here represent the principle components of the deviations among speakers from the speaker-mean MLLR supervector due to the speaker-specific characteristics. Each eigenvalue in $\Lambda_{\mathbf{w}}$ represents the magnitude of the variance along the direction of the corresponding eigenvector among the many speakers. Consequently, if we pick up properly top n eigenvectors $\{v_j, j = 1, 2, ..., n\}$ with relatively large eigenvalues, we may form an optimal subspace of the MLLR supervector space for speaker adaptation (as well as identification). The selection of the subspace is illustrated as a mask matrix Q_n in the following approximation formula 2.

$$C_{\mathbf{w}} \quad \cong \quad \left(E_{\mathbf{w}} Q_n\right) \, \Lambda_{\mathbf{w}} \, \left(E_{\mathbf{w}} Q_n\right)^{\mathrm{T}} \tag{2}$$

where Q_n is a mask matrix with 1 in the top n diagonal elements and 0 elsewhere.

2.3. Eigen-MLLR coefficients (EMCs) extracted by projection method

In view of the fact that the eigen-MLLR approach has been shown to be efficient in speaker adaptation [5][6], it is expectable that the coefficients in the eigen-MLLR space apparently carry good information for speaker variation. It is therefore proposed in this paper to use these parameters for speaker recognition. These parameters may contain much less irrelevant phonemic information than the MFCC features do. The projection method can be used here to compute one set of eigen-MLLR coefficients (EMCs) for each speech signal frame,

$$(c_j)_s \quad = \quad \left(o(t) - \bar{W}_s \, \xi_{s_r}\right)^{\mathrm{T}} \left((V_j)_s \, \xi_{s_r}\right) \tag{3}$$

where $(c_j)_s$, $j = 1..n$, are the EMCs to be estimated for class s, $o(t)$ the observed MFCC feature at time t, $\bar{W}_s$ the $D \times (D+1)$ MLLR matrix for the speaker-mean for class s, ξ_{s_r} the extended mean vector of mixture density s_r in the speaker-independent phonetic-acoustic model where the mixture density s_r belonging to the class s in the MLLR adaptation, and $(V_j)_s$ a $D \times (D+1)$ matrix for class s of the following form:

$$(V_j)_s \quad = \quad [(v_j)_s]_{D \times (D+1)}. \tag{4}$$

where $(v_j)_s$ is the j-th eigenvector for class s.

This frame-based projection method allows us to compare different size of training and test data sets flexibly, as will be clear in section 3. These coefficients definitely carry very efficient speaker-specific characteristics, because they represent the components of a vector for a specific speaker along the directions of eigenvectors of the eigen-MLLR space. This will be verified by the experiments below.

3. Experimental results

3.1. Experimental setup

The experiments were conducted on a PC dictation database of Mandarin Chinese recorded in Taiwan. Abundant speech data set (*A*) from 241 training speakers was used to train a set of speaker-independent phonetic-acoustic models, and to construct 240 eigenvectors based on the 241 MLLR matrices, where single-class MLLR is used. Up to ten seconds of GMM-training speech set (*B1*) produced by each of another 60 test speakers (30 female and 30 male) was then used in GMM training. The speaker-identification test was performed with another set (*B2*) of about 78 seconds of speech data in average produced by each of these 60 test speakers. In the tests, each decision was made by 5 seconds of observation window, with window shift of 4 seconds in test set. Table 1 summarizes the corpus.

data set	A	B1	B2
usage	SI model & eigen-MLLR	speaker-identification GMM training	GMM test
#speakers	241	60 (30 female, 30 male)	
size	72.5 hr.	10 sec. × 60	78 sec. × 60

Table 1: *Corpus description*

3.2. System configuration

The system consisted of a front-end speech signal preprocessor that converts speech utterance from its digital waveform representation into a stream of feature vectors, followed by a back-end speaker identifier that performs stochastic matching as well as decision making. A speaker-independent speech recognizer [8] was used at the front to generate the best state and mixture density sequence $\{ s_r\}$ for the speech signal in order to compute EMCs via Equation 3. During training, a group of L ($L = 60$) speakers $S = S_1, S_2, \ldots, S_L$ was represented by GMMs $\lambda_1, \lambda_2, \ldots, \lambda_L$ using feature vectors, in terms of either MFCCs or EMCs, extracted from the training data. Parameters of the GMMs were estimated via expectation-maximization algorithm [9]. In the testing phase, the system took as input the feature vectors extracted from the test utterance, and produced as output the likelihood score for each of the speaker models. According to the maximum likelihood decision rule, the identifier decided in favor of a speaker satisfying

$$\hat{S} = \underset{1 \le i \le L}{\operatorname{argmax}} \, p(O|\lambda_i). \tag{5}$$

where O is the sequence of feature vectors $o(t)$.

3.3. Baseline system — MFCC-based GMM

For performance comparison, a baseline system built upon MFCC-based features was evaluated first. A 22-dimension

Training data size	#Gaussian 8	16	32	64
10 sec.	12.46	6.27	**5.05**	7.08
7 sec.	18.24	12.87	**11.97**	16.37
5 sec.	24.92	**20.93**	25.00	37.54
3 sec.	**40.80**	41.94	57.25	80.78

Table 2: *Speaker identification error rate (%) for MFCC-based GMM*

feature vector consisted of 12 MFCCs plus 10 delta-MFCCs was extracted for every 25-ms of Hamming-windowed frame with 10-ms frame shifts. The speaker-identification performance with respect to the number of mixture components used in each speaker model are summarized in Table 2. As expected, the performance of MFCC-based GMMs depended highly on the quantity of the enrollment data. With less than 10 seconds of speech for each speaker, the performance dropped down rapidly.

3.4. EMC-based GMM

A series of preliminary experiments was first performed to choose the dimension n of the eigen-MLLR space, or the number of eigenvectors. These are speaker-identification tests using four-mixtured GMMs with 5 seconds of enrollment data for each speaker. From the results summarized in Table 3, we chose $n = 50$, or 50 eigenvectors, for the rest of experiments in the paper. After selecting the fifty dimensions for EMC features, the performance of speaker-identification tests for different number of mixture-components of GMM are shown partly in Table 4. The degradation of performance with enrollment data less than 10 seconds is apparently slower than that of MFCC-based GMM, and the performance is significantly better. The best performances of EMC features were all obtained with two mixture-components only, as compared to MFCCs' obtained with much higher mixture-components from 8 to 32.

n	22	30	50	70	100
error rate (%)	19.38	17.10	**12.14**	**12.13**	53.26

Table 3: *Selection of the dimension n of the eigen-MLLR space by speaker-identification tests with 4 mixtures of GMM and 5 seconds of enrollment data*

Training data size	#Gaussian 1	2	3
10 sec.	9.77	**5.54**	6.84
7 sec.	12.62	**8.22**	10.67
5 sec.	15.72	**10.91**	13.03
3 sec.	24.76	**19.30**	23.53

Table 4: *Speaker identification error rate (%) for EMC-based GMM*

3.5. MFCC/EMC-based joint GMM

The conventional MFCC features and the EMC features have quite different properties. It was observed in the experiments that these two types of features in fact perform differently for different speakers. A specific speaker might obtain very low accuracy with MFCC features while performed very well with EMC features, and vice versa. It is therefore possible to enhance the identification performance by properly combining these two different types of features together in the test. In such an MFCC/EMC-based joint GMM test, the probability evaluated for each frame of speech is simply

$$P\big(o_1(t), o_2(t)\big) = P_1^{(1-\alpha)}\big(o_1(t)\big) \cdot P_2^{\alpha}\big(o_2(t)\big) \quad (6)$$

where $o_1(t)$ and $o_2(t)$ are the MFCC and EMC feature vectors respectively, $P_1()$ and $P_2()$ are the MFCC-based and EMC-based GMM probabilities respectively, and α is a weighting factor, $0 \le \alpha \le 1$. The results in Table 3 show significant improvements of 50%-60% error rate reduction by this MFCC/EMC-based joint GMMs as compared to the conventional MFCC-based GMMs, regardless of the size of the training data being 3, 5, 7 or 10 seconds. In each case the achieved error rate is significantly lower than using either MFCC alone or EMC alone. The two types of feature parameters are apparently highly additive, which means the EMC features do bring extra speaker-specific information which are not present in MFCC features.

4. Further discussions

4.1. Analyses of discriminating functions for MFCC and EMC features

To analyze the discriminating functions of the two types of features, we evaluate the ratio of inter-speaker to intra-speaker variances for each feature dimension d, or the F-ratio F_d, according to the Fisher criterion function [10], based on the tested database of the 60 test speakers.

$$F_d = \frac{\textit{variance of speaker means}}{\textit{average intra-speaker variance}}. \quad (7)$$

The F-ratios F_d of MFCCs are in the range of 0.006 to 0.028 for the instantaneous MFCC components, at dimensions 1-12, while nearly 0 for the delta-MFCCs at dimensions 13-22, with an average of 0.0084, as shown in Figure 1 as a function of the 22 dimensions. The F-ratios of EMCs, on the other hand, are in the range of 0.03 to 0.25 at dimensions 1-50, with an average of 0.0784, roughly 10-times larger compared to that of MFCCs, as shown in Figure 2 as a function of the 50 dimensions of eigenspace. The general trend of smaller F_d ratio for higher dimensions is also clear in Figure 2.

4.2. Computation complexity analysis

It is worthwhile to compare the computation complexity required for using the proposed EMC features to that using the MFCC features. Consider first the memory requirement. A GMM with diagonal covariance matrices contains $(2D + 1)M$ free parameters, where D is the dimension of the features, and M is the number of mixture components. Therefore, with the above example system, we should store at least $(22 \times 2+1) \times 32 = 1440$ parameters when MFCC features are used, but $(50 \times 2 + 1) \times 2 = 202$ parameters when the proposed EMC features are used. Next consider the computational cost, we assume that multiplication, division, exponentiation, and logarithm all take a single multiply-add time, then a GMM-based system requires around $3DMTL$ multiply-add operations

Training data size	*EMC-based GMM* error rate (%)	#Gaussian	*MFCC/EMC-based joint GMM* error rate (%)	α	error rate reduction v.s. *MFCC*	*MFCC-based GMM* error rate (%)	#Gaussian
10 sec.	5.54	2	**2.52**	0.5	**50%**	5.05	32
7 sec.	8.22	2	**5.46**	0.7	**54%**	11.97	32
5 sec.	10.91	2	**8.39**	0.6	**60%**	20.93	16
3 sec.	19.30	2	**16.61**	0.7	**59%**	40.80	8

Table 3: *Speaker identification error rate (%) for MFCC/EMC-based joint GMM*

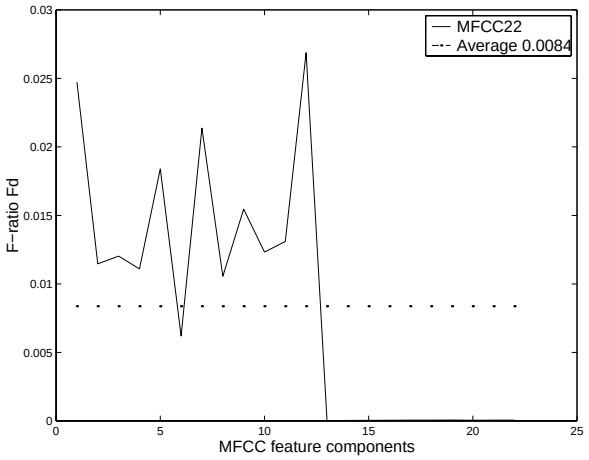

Figure 1: *F-ratio F_d of MFCCs over 60 speakers for the 22 dimensions*

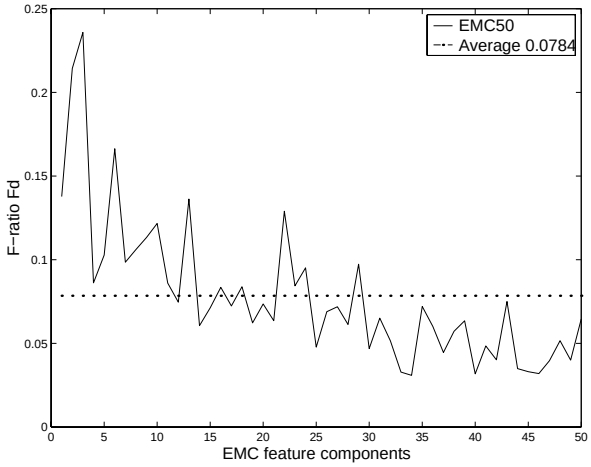

Figure 2: *F-ratio F_d of EMCs over 60 speakers for the 50 dimensions of eigenspce*

during the testing phase, where T is the utterance length and L is the number of speakers. Therefore, the number of multiply-add operations involving MFCC and EMC features for the present system are $3 \times 22 \times 32TL = 2112TL$ for MFCCs and $3 \times 50 \times 2TL = 300TL$ for EMCs, respectively. The extra computation required for the EMC extraction in Equation 3 is very limited. If the application system includes a speech recognizer, the additional cost in state and mixture density alignment is also limited. It is thus clear that the proposed EMC-based GMMs require much less memory and computational cost, compared to the MFCC-based GMMs.

5. Conclusions

Eigen-MLLR Coefficient (EMC) features are proposed for speaker identification in this paper. The F-ratios of EMC features were shown in average about 10-times larger than those of MFCC features. EMC features outperformed MFCC features significantly, especially with limited training data. Using both MFCC and EMC features in a joint GMM can achieve error rate reduction of 50%-60% as compared to using only MFCC features. It is worth noting that the proposed method can create an advantageous scenario that integrating speech and speaker recognition to facilitate the goal of human-machine communication.

6. Acknowledgements

We would like to thank Sammy Lee for his help on eigen-MLLR training.

7. References

[1] D.A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. In *Speech Communication*, v. 17, pp. 177-192, 1995.

[2] O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua Speaker identification and verification using eigenvoices In *Proc. ICSLP*, v. 2, pp. 242-245, Beijing, 2000.

[3] C.J. Leggetter & P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In *Computer Speech & Language*, v. 9, pp. 171–185, 1995.

[4] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, & M. Contolini. eigenvoices for speaker adaptation. In *Proc. ICSLP*, v. 5, pp. 1771–1774, Sydney, 1998.

[5] K.-T. Chen, W.-W. Liau, H.-H. Wang, & L.-S. Lee. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression. In *Proc. ICSLP*, v. 3, pp. 742-745, Beijing, 2000.

[6] N. Wang, S. Lee, F. Seide, & L.-S. Lee. Rapid speaker adaptation using A priori knowledge by eigenspace analysis of MLLR parameters. To appear in *Proc. ICASSP*, Salt Lake City, 2001.

[7] I.T. Jolliffe. Principle component analysis. Springer, 1986.

[8] F. Seide and N. Wang. Phonetic modelling in the Philips Chinese continuous-speech recognition system. In *Proc. ISCSLP*, pp. 54–59, Singapore, 1998.

[9] A.P. Dempster, N.M. Laird, & D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm In *Journal of the Royal Statistical Society*, ser. B, v. 39, pp. 1–38, 1997.

[10] R.O. Duda & P.E. Hart. Pattern classification and scene analysis. John Wiley & Sons, Inc., 1973.