Language Identification Using Acoustic Models an Speaker Compensated Cepstral-Time Matrices

LANGUAGE IDENTIFICATION USING ACOUSTIC MODELS AND SPEAKER COMPENSATED CEPSTRAL-TIME MATRICES

Fabio Castaldo^, Emanuele Dalmasso^, Pietro Laface^ , Daniele Colibro, Claudio Vair**
Politecnico di Torino, Italy^
{Fabio.Castaldo,Emanuele.Dalmasso,Pietro.Laface}@polito.it
Loquendo, Torino , Italy*
{Daniele.Colibro,Claudio.Vair}@loquendo.com

ABSTRACT

This work presents two contributions to language identification. The first contribution is the definition of a set of properly selected time-frequency features that are a valid alternative to the commonly used Shifted Delta Cepstral features.

As a second contribution, we show that significant performance improvement in language recognition can be obtained estimating a subspace that represents the distortions due to inter-speaker variability within the same language, and compensating these distortions in the domain of the features.

Experiments on the NIST 1996 and 2003 Language Recognition Evaluation data have been successfully used to validate the effectiveness of the proposed techniques.

Index Terms— language identification, cepstral-time matrices, speaker and channel compensation, frame domain compensation, phonetic language models

1. INTRODUCTION

The combination of acoustic based Language Identification (LID) systems with phonetic systems has been shown to give excellent performances in the last formal NIST evaluations [1,2]. This paper focuses on acoustic only LID systems for which Gaussian Mixture Modeling (GMM) and Support Vector Machine (SVM) are the state-of-the-art classifiers [3,4]. The main advantage of acoustic-based systems is that they do not require phonetic transcriptions. Moreover, improved acoustic systems allow obtaining better performance in combination with phonetic-based systems.

In this paper we present the cepstral-time matrices [5] as an alternative to the commonly used Shifted Delta Cepstra (SDC) features. These time-frequency features in our advice have more perceptual grounds and wider flexibility. Moreover, using less parameters, they give results similar to the SDC features.

Also, to reduce inter-speaker variability rather than using Vocal Tract Length Normalization [6,2], we show that significant performance improvement can be obtained using factor analysis. In particular, we evaluate a factor subspace that represents the distortions due to inter-speaker variability within the same language, and compensate these distortions in the domain of the features.

The paper is organized as follows: Section 2 analyzes the characteristics of the SDC features and allows appreciating another approach for capturing temporal dependencies, described in Section 3, based on cepstral-time matrices. An extensive set of experiments using these features with a Support Vector Machine classifier is presented in Section 4. Section 5 and 6 present the frame based inter-speaker variation compensation approach and its performance, respectively. Last Section is devoted to our final remarks and ongoing work.

2. SHIFTED DELTA FEATURES

The Shifted Delta Cepstral features have been introduced to improve the LID performance with respect to the classical cepstral and delta cepstral features [7].

The SDC coefficients are computed, for a cepstral frame at time *t*, according to:

$$\Delta c_n(t,i) = c_n(t+iP+d) - c_n(t+iP-d) \qquad n = 0, N-1 \qquad i = 0, k-1 \tag{1}$$

where *n* is the *n-th* cepstral coefficients, *d* is the lag of the deltas, *P* is the distance between successive delta computations, and *i =0, k-1* is the SDC block number. The final feature vector is obtained by concatenation of *k* blocks of *N* parameters. The configuration 7-1-3-7 for N-d-P-k has been used for language recognition in [4] in the framework of the generalized linear discriminant sequence kernel (GLDS) Support Vector Machines (SVM) [8]. In this framework, the mean SDC vector **b** is computed averaging the SDC coefficients of all the utterance frames.

3. CEPSTRAL-TIME MATRICES

The main rational for using SDC is to incorporate additional temporal information about the speech into the feature vector [7] to capture temporal dependencies that are typical of a language.

For this purpose, time-frequency features, have been used since long time for speech recognition [5,10,11], and more recently for speaker [12], and language recognition [13].

The time-frequency features in the approach presented in [13] require the estimation of one "filter", obtained by Principal Component Analysis, for each language. Moreover they have been tested only for a relatively small context of 9 frames, corresponding to the classical delta-delta cepstral window.

Cepstral-time matrices [5] account both for short and long time variations of the spectral features and their correlation in time. A single one-dimensional DCT along the time axis of a matrix containing *W* MFCC vectors (the context window) is required to produce the cepstral-time matrices, rather than several PCA filters as in the time-frequency approach of [13]. Moreover, the number of the cepstral coefficients, the order of the temporal DCT, and the length of the context window can be properly selected to capture different types of spectral variations, and better performance can be obtained combining different systems exploiting the complementarities of these features.

Table 1. %EER of an SVM classifier with different features using the best four sets of features.

Corpus	N. of params	Features	SVM without frame compensation		
			30sec	10sec	3sec
NIST LRE 1996	49	SDC 7-1-3-7 *No MFCC*	4.38	12.55	25.98
	56	SDC 7-1-3-7	3.11	9.98	23.31
	48	DCT 12-3-7	3.66	9.10	21.63
	49	DCT 7-6-21	3.25	10.51	23.38
	50	DCT 5-9-21	3.11	11.13	23.65
NIST LRE 2003	49	SDC 7-1-3-7 *No MFCC*	6.59	14.74	27.02
	56	SDC 7-1-3-7	4.43	12.06	23.77
	48	DCT 12-3-7	4.85	11.97	23.68
	49	DCT 7-6-21	4.60	12.47	25.69
	50	DCT 5-9-21	5.60	14.22	27.44

4. COMPARING SDC AND CEPTRAL-TIME FEATURES

In this work we tested three cepstral-time feature configurations where *static cepstral coefficients* are concatenated with the parameters of a *cepstral-time matrix* obtained by a temporal DCT. Settings referred to as DCT *N-O-W* define a cepstral-time matrix where a temporal DCT of order *O* is performed on a context window of *W* frames including the first *N* MFCCs (*0* to *N-1*). In particular, we tested the settings:

- DCT 12-3-7, which produces information related to the first, second, and third order differentials of the first 12 MFCCs
- DCT 7-6-21, which has the same number of parameters of the SCD 7-1-3-7 features, and covers long time variations of the spectral envelope. A context of 21 frames approximately corresponds to the duration of a syllable.
- DCT 5-9-21 that has almost the same number of parameters of the previous ones, but covers a different area in the frequency-quenfrency plane of the cepstral-time matrix excluding long time variations of the pitch and introducing some short time variations of the spectral envelope [5].

We did not investigate extensively the best settings for the DCT parameters, focusing only on the ones that provide the same number of parameters, and/or have similar characteristics with respect to the commonly used features.

All the experiments aiming at comparing different sets of features have been performed on the NIST 1996 and 2003 Language Recognition Evaluation (LRE) data [14-15], and according to NIST evaluation rules.

The test corpora include 12 target languages: American English, Arabic, Canadian French, Farsi, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese. Russian has been used as the out-of-language in the 2003 tests. In these evaluations there are three duration settings, 3, 10, and 30 seconds. The 1996 evaluation data consist of 1,503, 1,501 and 1,492 sessions of 3, 10, and 30 seconds, respectively. The 2003 evaluation data consists of 1,280 sessions for each duration setting.

A gender independent model per language has been created using the training and development sets of the CALLFRIEND corpus [15]. For training, we segmented each conversation in this corpus into ~150 sec sessions.

Table 2. %EER of the score combination of two SVM classifiers using different features on the 30sec tests of the NIST LRE 1996 (upper right) and NIST LRE 03 (lower left).

Features	SDC 7-1-3-7	DCT 7-6-21	DCT 5-9-21	DCT 12-3-7
SDC 7-1-3-7		2.62	2.28	2.08
DCT 7-6-21	3.70		2.43	2.15
DCT 5-9-21	3.51	3.42		**1.90**
DCT 12-3-7	**3.09**	3.25	**3.09**	

The GLDS SVM approach of [4], with a slight different normalization of the expanded vector **b,** has been used in the first set of experiments. The SVM scores were converted to log-likelihood ratios [4].

The results of our experiments are given in terms of Equal Error Rate (EER) defined as the error of the language detection system when the detection threshold is set so that the probability of false alarms equals the probability of misses. The prior probability is uniform among the languages. The results, given in Table 1, confirm that adding the static cepstral parameters to the SDC feature vector sensibly improves the recognition performance [2].

The DCT 7-6-21 features (42+7 static parameters) give comparable results with respect to the SDC 7-1-3-7 (49+7 parameters), while the DCT 12-3-7 features are the best for the 10sec and 3sec tests.

Since it is well known that combining different sources of information typically improves LID performance, experiments have been performed with the pairwise combination of systems using the best four sets of features. The results obtained on the NIST LRE 2003 30sec tests are shown in lower left part of Table 2. The weights of the Neural Network used for the score combination have been trained on the test set of the CALLFRIEND corpus.

The combination of either SDC 7-1-3-7 or DCT 7-6-21 with DCT 12-3-7 – the most complementary set of features – gives similar results. It is also worth noting that, although the DCT 5-9-21 features have the worst performance (see Table 1), their combination with DCT 12-3-7 achieve the best equal error rate, demonstrating their complementarities. The same considerations remain valid examining the upper right part of Table 2, showing the results on the NIST LRE 1996 30sec test data. The 10sec and 3sec tests confirm these results.

5. INTER-SPEAKER VARIABILITY COMPENSATION

In language recognition, errors are due not only to the similarity among the models of different languages, but also to the intrinsic variability of different utterances of the same language. The performance is heavily affected when a model, trained in a set of conditions, is used to test data collected from different speakers, microphones, channels, and environments. In this paper we will refer to all these mismatching conditions as inter-speaker variability.

Model-based techniques have been recently proposed for *speaker recognition*, which are able to compensate speaker and channel variations without requiring the explicit identification and labeling of different conditions. These techniques share a common background: modeling the variability of speaker

utterances constraining them to a low dimensional space [16-18]. In [19] we have proposed a solution for *speaker recognition* in the GMM framework that allows compensating the observation features rather than models parameters. Compensating features rather than models has the advantage that the transformed parameters can be used as observation vectors for classifiers of different nature and complexity. In this Section we recall the main steps of this approach for compensating the speaker intersession variability, and we show how it can be used as well for compensating inter-speaker variability within a language.

In most state-of-the-art approaches, the speaker models are derived from a common GMM root model, the so called Universal Background Model (UBM), by means of MAP adaptation [1]. A supervector that includes all the speaker specific parameters can be obtained simply appending the adapted mean value of all the Gaussians in a single stream. The same can be done for the UBM, obtaining the UBM supervector. The distortions in the large supervector space can be summarized by a small number of parameters - the *channel factors* [18] -, in a lower dimensional subspace.

5.1 Model-domain adaptation

Channel factors adaptation for an utterance i and a supervector k is performed, in the supervector model space, as follows:

$$\boldsymbol{\mu}^{(i,k)} = \boldsymbol{\mu}^{(k)} + \mathbf{U}\mathbf{x}^{(i,k)} \qquad (1)$$

where $\boldsymbol{\mu}^{(i,k)}$ and $\boldsymbol{\mu}^{(k)}$ are the adapted and the original supervector of GMM k respectively. $\mathbf{U}$ is a low rank matrix projecting the channel factors subspace in the supervector domain. The N-dimensional vector $\mathbf{x}^{(i,k)}$ holds the channel factors for the current utterance i and GMM k. The $\boldsymbol{\mu}^{(k)}$ supervectors are obtained by the classical MAP adaptation

We have shown in [19] that since the vector $\mathbf{x}^{(i,k)}$ accounts for the distortions produced in the supervector space by the intersession variability, it depends on the utterance i, but only weakly on the speaker model k. Thus, it can be estimated using the UBM, i.e. dropping the dependence on the GMM k. This is equivalent to set $\mathbf{x}^{(i,k)} = \mathbf{x}^{(i)} \;\; \forall k$ and to apply the normalization:

$$\boldsymbol{\mu}^{(i,k)} = \boldsymbol{\mu}^{(k)} + \mathbf{U}\mathbf{x}^{(i)} \qquad (2)$$

for all the models k that must be scored against utterance i.
For each test utterance i in equation (2), we estimate vector $\mathbf{x}^{(i)}$ with a single iteration of a technique called Probabilistic Subspace Adaptation (PSA) [20].

5.2 Estimation of the subspace U

For *speaker recognition*, the channel factors subspace, modeled by the low rank matrix $\mathbf{U}$, is assumed to represent the distortion due to the intersession variability. This distortion can be estimated analyzing how the models of the *same speaker* are affected, when trained with utterances collected from different channels or conditions. Thus, the $\mathbf{U}$ matrix is estimated with a large set of differences between models, obtained by MAP adaptation, using different utterances of the same speaker.

For *language recognition* we are interested, instead, in compensating the distortions due to inter-speaker variability within the *same language*. Thus, the $\mathbf{U}$ matrix is estimated with a large set of differences between models generated using different speaker utterances of the same language.

Table 3. %EER Effects of inter-speaker variations compensation using GMM classifiers with DCT 7-6-21

Corpus	GMM without frame compensation			GMM with frame compensation		
	30sec	10sec	3sec	30sec	10sec	3sec
NIST 1996	6.92	9.87	19.00	2.35	6.67	17.08
NIST 2003	9.25	12.5	20.16	4.08	8.07	18.17

5.3 Feature-domain adaptation

Relying on the hypotheses that led to equation (2), we assume that the acoustic space distortion, characterized by the vector $\mathbf{x}^{(i)}$, can be estimated using the UBM rather than the speaker dependent model GMM k. Neglecting, for the sake of conciseness, the model index k, we rewrite (2) for each Gaussian component m of the supervector as:

$$\boldsymbol{\mu}_m^{(i)} = \boldsymbol{\mu}_m + \mathbf{U}_m\mathbf{x}^{(i)} \qquad \forall m \qquad (3)$$

where of $\boldsymbol{\mu}_m^{(i)}$, $\boldsymbol{\mu}_m$ and $\mathbf{U}_m$ all refer to the m-th Gaussian of the GMM. The number of rows of the mean vectors and of the subspace matrix $\mathbf{U}_m$, is equal to the dimension of the input feature vector.

The adaptation of the feature vector at time frame t, $\mathbf{O}(t)$, is obtained by subtracting to the observation feature a weighted sum of the channel compensation offset values:

$$\hat{\mathbf{O}}^{(i)}(t) = \mathbf{O}^{(i)}(t) - \sum_m \gamma_m(t)\,\mathbf{U}_m\mathbf{x}^{(i)} \qquad (4)$$

where $\gamma_m(t)$ is the Gaussian occupation probability, and $\mathbf{U}_m\,\mathbf{x}^{(i)}$ is the channel compensation offset related to the m-th Gaussian of the UBM model. In the actual implementation, the right side summation of (4) is limited, for the sake of efficiency, to the first best contributes only.

Our feature adaptation approach has shown in [19] to give the same benefits of the model domain adaptation. A system, based on this technique, was among the best participating to the NIST 2006 Speaker Recognition Evaluation.

6. EXPERIMENTS USING GMM CLASSIFIERS

The subspace $\mathbf{U}$ matrix and two 512 Gaussian gender dependent UBMs have been trained using the training and development sets of the CALLFRIEND corpus [15]. A gender dependent model for each language in the NIST 1996 and 2003 LRE has been MAP adapted using the same data. The use of the UBM allows not only to speed-up both training and testing, but also to perform frame compensation for reducing inter-speaker variability.

During testing, the UBM gender model that produces the best likelihood is selected together with the set of its corresponding gender dependent language models. The final score for each language includes both T-normalization and log-likelihood ratio normalization.

Table 4. %EER of a GMM classifier using different features

Corpus	N. of param.	Features	GMM with frame compensation		
			30sec	10sec	3sec
NIST 1996	56	SDC 7-1-3-7	1.88	5.65	15.48
	49	DCT 7-6-21	2.35	6.67	17.08
	48	DCT 12-3-7	4.43	7.73	16.38
	50	DCT 5-9-21	2.75	7.40	19.16
NIST 2003	56	SDC 7-1-3-7	3.67	8,16	17,26
	49	DCT 7-6-21	4.08	8.07	18.17
	48	DCT 12-3-7	6.33	10.25	19.34
	50	DCT 5-9-21	4.83	10.00	19.42

Table 5. %EER of the score combination of GMM and SVM classifier using different features (NIST 2003, 30sec tests)

CLASSIFIERS		GMM	SVM			
	Features	SDC 7-1-3-7	SDC 7-1-3-7	DCT 7-6-21	DCT 5-9-21	DCT 12-3-7
GMM	SDC 7-1-3-7	-	**2.52**	2.59	2.58	**2.50**
	DCT 7-6-21	3.41	2.75	2.66	**2.50**	2.84
	DCT 5-9-21	3.25	2.94	2.85	2.76	3.26
	DCT 12-3-7	3.83	3.16	3.76	3.02	3.40

Table 3 shows that the compensation of the inter-speaker variations provides relevant performance improvement, increasing with the length of the utterance. The comparison has been done for the DCT 7-6-21 features, which are the most similar to the SDC 7-1-3-7 ones.

Even for the GMM classifiers, a comparison of the performance on the NIST 2003 LRE tests using different features, reported in Table 4, confirm that the DCT 7-6-21 parameters are a good alternative to the SDC ones. Moreover, while the combination of the scores obtained with GMMs using different features reduces only slightly the equal error rate, as shown in the first column of Table 5, a significant performance increase is obtained by combining GMM and SVM scores. Also, there are two combinations of DCT features give the same results of the combination of two SDC-based classifiers.

7. CONCLUSIONS

We have presented the cepstral-time features, an alternative to Shifted Delta Cepstra that, in our advice, has more perceptual grounds, has wider flexibility, and give similar results to MFCC+SDC with less parameters. We have shown that classifiers using sets of DCT features that are "orthogonal" in the frequency-quefrency plane present good complementarities, as shown by the performance increase obtained with their combination. The results applying inter-speaker variations compensation confirmed the quality of the frame compensation approach already assessed on speaker recognition experiments.

The reported results are well aligned with the best ones recently reported on the NIST LRE 2003 task [21] with more complex systems [1], and with the ones obtained *without discriminative training* in [2].

Work is in progress on the combination of this acoustic approach (possibly discriminatively trained) with a phonetic one, where the phonemes of an utterance are estimated by the Loquendo ASR recognizer.

8. REFERENCES

[1] W. Campbell, T. Gleason, J. Navratil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced Language Recognition Using Cepstra and Phonotactics: MITLL System Performance on the NIST 2005 Language Recognition Evaluation," IEEE Odyssey Speaker and Language Recognition Workshop, Puerto Rico, 2006.

[2] L. Burget, P. Matejka, and J. Cernocky, "Discriminative Training Techniques for Acoustic Language Identification," Proc. ICASSP 2006, Vol I, pp. 209-212, 2006.

[3] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[4] W.M. Campbell, E. Singer, P.A. Torres-Carrasquillo, and D.A. Reynolds, "Language Recognition with Support Vector Machines," Proc. Odyssey: The Speaker and Language Recognition Workshop, ISCA, pp. 41-44, 2004.

[5] B.P. Milner, and S.V. Vaseghi, "An Analysis of Cepstral-Time Matrices for Noise and Channel Robust Speech Recognition," Proc. EUROSPEECH 1995, pp. 519-522, 1995.

[6] E. Wong, and S. Sridharan, "Methods to Improve Gaussian Mixture Model Based Language Identification Systems," Proc. ICSLP 2002, pp. 93-96, 2002.

[7] P.A. Torres-Carrasquillo, D.A. Reynolds, E. Singer, M.A. Kohler, R.J. Greene, and J.R. Deller, Jr., "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," Proc. ICSLP 2002, pp. 89-92, 2002.

[8] W.M. Campbell, "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition," Proc. ICASSP 2002, pp. 161-164, 2002.

[10] L. Fissore, P. Laface, and F. Ravera, "Using word temporal structure in HMM speech recognition," Proc. ICASSP 1997, Vol II, pp. 975-978, 1997.

[11] D. Macho, C. Nadeu, P. Jancovic, G. Rozinaj, and J. Hernando, "Comparison of Time & Frequency Filtering and Cepstral-Time Matrix Approaches in ASR", Eurospeech 1999, pp. 77-80, 1999.

[12] I. Magrin-Chagnolleau, G. Durou, and F. Bimbot, "Application of Time-Frequency Principal Component Analysis to Text-Independent Speaker Identification," IEEE Trans. On Speech and Audio Processing, Vol. 10, n. 6, pp. 371-378, 2002.

[13] M. Dutat, I. Magrin-Chagnolleau, and F. Bimbot., "Language Recognition using Time-Frequency Principal Component Analysis and Acoustic Modeling," Proc. ICSLP 2000, Vol.II, pp. 230-233, 2000.

[14] http://www.nist.gov/speech/tests/lang/index.htm

[15] http://www.ldc.upenn.edu/Catalog/

[16] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling with Sparse Training Data", IEEE Trans. on Speech and Audio Processing, Vol.13, No.3, pp. 345-354, May 2005.

[17] R. Vogt, B. Baker, and S. Sridharan, "Modelling Session Variability in Text-Independent Speaker Verification," Proc. INTERSPEECH 2005, pp. 3117-3120, 2005.

[18] P. Kenny, and P. Dumouchel, "Disentangling Speaker and Channel Effects in Speaker Verification," Proc. ICASSP 2004, Vol. I, pp. 37-40, 2004.

[19] C. Vair, D Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel Factors Compensation in Model and Feature Domain for Speaker Recognition," IEEE Odyssey Speaker and Language Recognition Workshop, Puerto Rico, 2006.

[20] S. Lucey, and T. Chen, "Improved Speaker Verification Through Probabilistic Subspace Adaptation," Proc. EUROSPEECH 2003, pp. 2021-2024, 2003.

[21] A.F. Martin and M.A. Przybocki, "NIST 2003 language recognition evaluation," in Proc. EUROSPEECH 2003, pp. 1341–1344, 2003.