# A FURTHER INVESTIGATION ON SPEECH FEATURES FOR SPEAKER CHARACTERIZATION

*I. Magrin-Chagnolleau, G. Gravier, M. Seck, O. Boeffard*[*]*, R. Blouet, F. Bimbot.*

IRISA (CNRS & INRIA), Campus universitaire de Beaulieu, 35042 Rennes cedex, France – European Union.

[*] *also with* ENSSAT, 6 rue de Kerampont, BP447, 22305 Lannion cedex, France – European Union.

ivan@ieee.org - ggravier@irisa.fr - mseck@irisa.fr

olivier.boeffard@enssat.fr - rblouet@irisa.fr - bimbot@irisa.fr

## ABSTRACT

In this article, we investigate on alternative speech features for speaker characterization. We study Line Spectrum Pairs features, Time-Frequency Principal Components and Discriminant Components of the Spectrum. These alternative features are tested and compared on a task of speaker verification. This task consists in verifying a claimed identity from a speech segment. Systems are evaluated on a subset of the evaluation data of the NIST 1999 speaker recognition campaign. The new speech features are also compared to the classical cepstral coefficients, which remain, in our experiments, the best performing features.

## 1. INTRODUCTION

Cepstral coefficients and their delta parameters have been intensively used in speech processing. Although they have been developped for speech recognition, they are also used for speaker recognition mainly because of their nice statistical properties and the possibility of doing channel compensation in the cepstral domain. In this paper, we investigate on alternative speech features, namely Line Spectrum Pairs (LSP), Time-Frequency Principal Components (TFPC) and Discriminant Components of the Spectrum (DCS).

LSP features have been used for speech coding [9] but rarely for speaker characterization. Liu *et al.* studied LSP derived parameters in a VQ based text-dependent speaker verification system and concluded towards a better performance of the LSP frequencies over the cepstral coefficients [1]. In this paper, we use these parameters for text-independent speaker verification using Gaussian mixture models (GMM).

The TFPC analysis [2] is a way of capturing dynamic information in the speech signal by a principal component analysis of the contextual matrix, which is the covariance matrix of feature vectors augmented by their time context. In former work, TFPC has shown interesting properties in enhancing some speaker-specific information.

DCS features are obtained by a Linear Discriminant Analysis on the log-magnitude spectrum. LDA has already been used for speech analysis in speaker verification. However, in our approach, we take into account a priori knowledge, such as the gender of the speaker and the handset type of the call, for estimating the discriminant components.

These three speech features are tested and compared on a speaker verification task. This task consists in verifying a claimed identity from a speech segment. The experiments are conducted on a subset of the evaluation data of the NIST 99 speaker recognition campaign. This corpus was extracted from the Switchboard corpus and is composed of conversations recorded over the telephone.

## 2. FEATURES

### 2.1. Line spectrum Pairs

LSP frequencies are related to linear predictive analysis. Considering the $p$ order LPC inverse filter $A_p(z)$, the LSP frequencies are defined as the zeros of the two polynoms defined by $A_p(z) \pm z^{-(p+1)}A_p(z^{-1})$. It can be shown that these zeros are on the unit circle and correspond to frequencies related to the formant frequencies of the LPC filter [9].

The $p$ LSP frequencies are ordered to form a feature vector. The main difficulty with the LSP comes from the fact that the features are ordered and take their values in a bounded interval, namely $]0, 0.5[$. These two properties are not very suited for Gaussian modeling. In order to overcome the ordering property, the LSP frequencies are centered by subtracting the long term mean of each feature. In this pa-

A part of this work has been done in the framework of the first phase of the AGIR project, funded by the French RNRT Programme.

per, 16th order linear prediction is considered. The frame length is set to 20 ms with a 50% overlap.

### 2.2. Time-Frequency Principal Components

A potential way to characterize a speaker is to extract time-frequency patterns that are characteristic of the acoustic vector trajectory of that speaker. In the case of TFPC, these time-frequency patterns are obtained by calculating the principal components of the contextual feature vector, which is the feature vector augmented by its time context. The original sequence must be long enough to be representative of the speaker we want to characterize with the time-frequency patterns.

Once the patterns have been extracted, they are used to filter the acoustic trajectory of the training material and of the test material. Any modeling technique can then be applied to the new vectors, as it is done usually on cepstral vectors. For a detailed presentation of the TFPC analysis, see [2, 3].

### 2.3. Discriminant Components of the Spectrum

In the case of DCS, we consider the set of spectral features (primary vectors) of different speakers as belonging to different classes. Linear Discriminant Analysis (LDA) aims at finding a new feature vector, obtained by linear transformation of the primary vectors, with lower dimension and which maximally separates the different classes. Here, a class corresponds to a speaker, and primary vectors are the log-module of the spectrum magnitude, obtained by Fourier Transform.

We learn the LDA on a variety of speakers, in order to obtain features which are speaker-discriminant on the learning set. Coefficients of the linear discriminant transformation are the eigenvectors of the matrix $V^{-1}B$, where $V$ is the total covariance matrix and $B$ is the covariance matrix between classes. Discriminant components are selected according to their corresponding eigenvalues (in decreasing order) and form the Discriminant Components of the Spectrum (DCS).

The result of the LDA depends on the composition of analysis data, between the different speakers and the two handset types[1]. For taking care of the gender variety, we use the same number of speakers for each gender. For handling the different type of handsets, we weight the electret and carbon observations, in order to have equal total weights for the two types of data. As a consequence, observation from speaker $s$ and handset type $h$, have the weight $\frac{1}{2*S*N_{sh}}$, where $S$ is the number of learning speakers, and $N_{sh}$ is the the number of observations from speaker $s$ and handset type $h$. In this experiment, $S = 50$.

Note that an important limitation of DCS, is the impossibility of applying mean subtraction (which is however a well-known technique for channel mismatch compensation). In fact, if mean subtraction were applied in the spectral domain, class means would all be equal to zero, leading to a null covariance matrix between classes $B$.

In this paper, each speaker have at least 30 seconds of data for each handset type. For robustness of LDA to channel effects, different recordings are also used for a speaker. Frames of 16 ms duration with 50% overlap are used. Spectral analysis is restricted to the telephone band [340 Hz, 3400 Hz]. Features are 16 DCS, augmented by their delta and the delta log-energy.

## 3. EXPERIMENTS

### 3.1. Database

The evaluation data for the NIST 1999 speaker recognition campaign [6] comes from the Switchboard 2–Phase 3 corpus collected by the Linguistic Data Consortium (LDC). The training data for each target speaker consist of two utterances of 1 minute each in average, obtained by concatenating consecutive turns of the speaker. Segments of silence were removed. We use a subset of this corpus for our experiments. The total number of target speakers is 119 (79 female speakers and 40 male speakers).

The test data consist of segments containing speech of only one speaker, obtained by concatenating consecutive turns of this speaker. Here also, segments of silence were removed. The duration of the test segments is between a few seconds and 1 min. There are 857 test segments, tested against several claimed identities, totaling 9427 tests. Claimed speakers have always the same gender than the test segment speaker.

Different test conditions are considered by NIST. Test conditions are defined according to the call phone number and handset type mismatches, between the test segment and train data (when the claimed identity is the good one). The four test conditions are : Same Number / Same Type (SNST), Same Number / Different Type (SNDT), Different Number / Same Type (DNST) and Different Number / Different Type (DNDT).

### 3.2. Verification system

The verification system is a variant of the ELISA Text-Independent Speaker Verification Platform [7]. It proceeds by

[1] In the USA, two types of handset are used : electret and carbon.

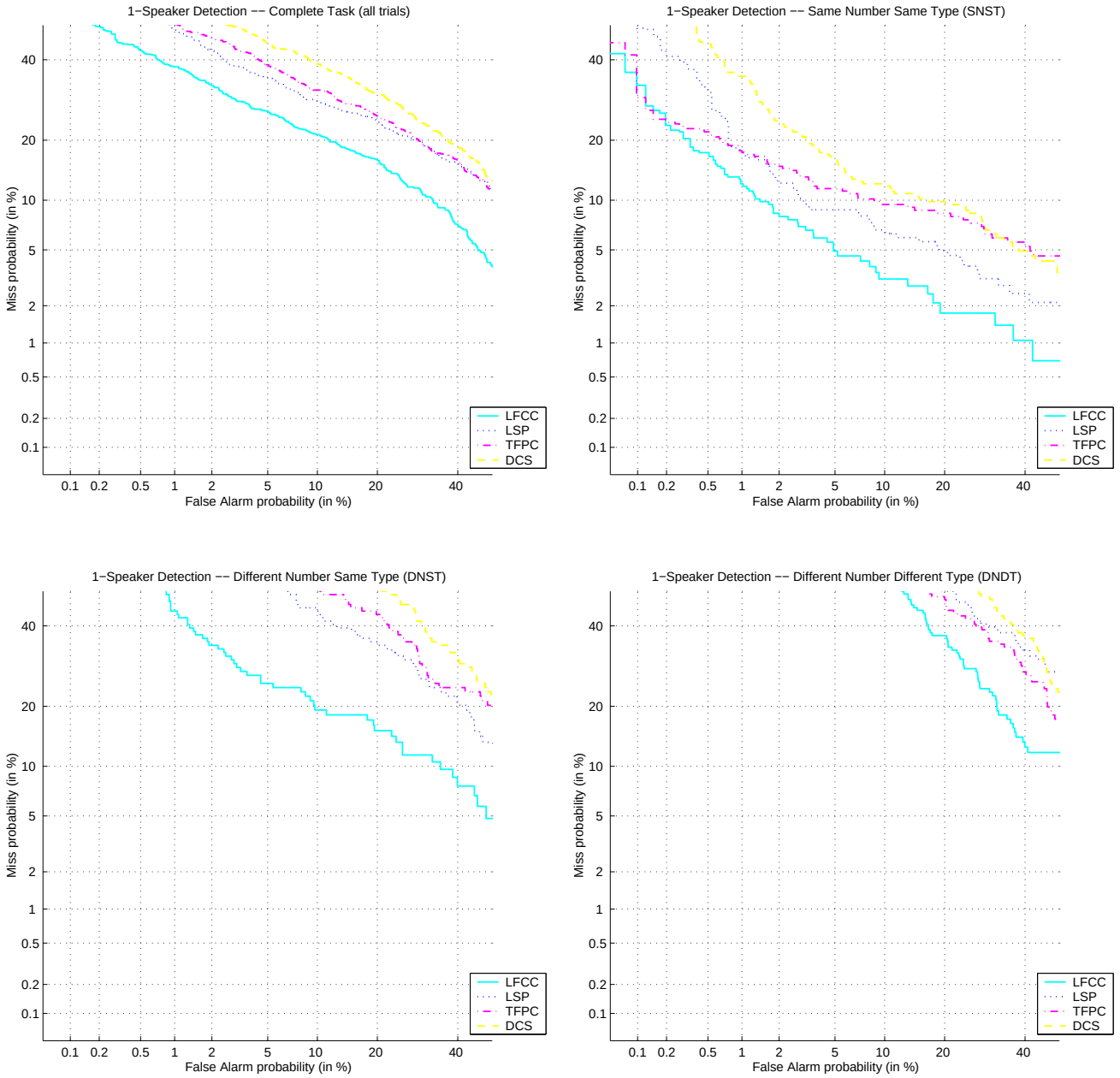


Figure 1: Performances of LFCC, LSP, TFPC and DCS, on all trials and for the three following NIST test conditions : SNST, DNST, DNDT

training a target speaker model for each client-speaker, and a speaker-independent world-model for each gender. Both speaker models and world models are mixtures of 128 Gaussians with diagonal covariance matrices [8]. For each frame in a test segment, the log-likelihood ratio (LLR) between the target speaker model and the wolrd-model is calculated. The utterance score is obtained by averaging the frame-based LLR. This score is compared to a threshold to make the final decision.

We compare the features defined in section 2 with a reference configuration based on Linear Frequency Cepstral Coefficients (LFCC) plus their first order deltas augmented with the delta log energy. For the reference feature set, a

24 channel triangular filter-bank is applied to 20 ms frames every 10 ms. The 24 filter central frequencies are regularly spread along the frequency range [340 Hz - 3400 Hz]. Finally, cepstral coefficients $c_1$ to $c_{16}$ are computed from the filter bank output.

### 3.3. Evaluation

The evaluation of the systems is done using Detection Error Trade off (DET) curves [4]. This representation is a way to show all of the possible operating points of a system (false alarm rate vs miss rate) in a scale which makes the result curves rather linear and easier to compare.

## 4. RESULTS

Figure 1 represents the performance of the different features, when evaluated on all trials and when restricted to the three following test conditions : SNST, DNST, DNDT.

As usually observed, performance decreases when the call phone number is different (from SNST to DNST) and drops further when the type of handset changes (DNST to DNDT). All four feature sets seem very sensitive to phone number (channel) mismatch effects.

The LSP and TFPC alternate in second and third position, according to the handset-type condition. This suggests that LSP may be more sensitive than TFPC to handset type mismatch.

DCS perform systematically worse than the other feature sets. This result may be explained by the fact that the population of learning speakers is rather limited in this experiment. Note also that, as mentioned earlier, no channel compensation was carried out.

Finally, all tested parameters are clearly outperformed by the conventional LFCC, which shows that, even though these acoustic features are not specifically designed for speaker characterization, they possess good properties for this task.

## 5. SUMMARY AND CONCLUSIONS

Our experiments have not succedeed in evidencing a benefit of alternate acoustic features over cepstral coefficients, for speaker characterization. This may be partly owed to the lack in implementing equivalent pre- and post-processing for computing the alternate parameters. Future efforts should focus on attempts in unbounding the LSP, normalizing the likelihood ratio according to the speaker-dependent transformation for TFPC, better addressing channel compensation issues for the DCS, etc...

However, our experiments confirm the difficulty in finding, for speaker verification, better acoustic features than the conventional cepstrum coefficients, for which the theory and the know-how are well mastered.

## 6. REFERENCES


[1] Chi-Shi Liu, Min-Tau Lin, Wern-Jun Wang and Hsiao-Chuan Wang Study of Line Spectrum Pair frequencies for speaker recognition. *ICASSP*, 1990.

[2] Ivan Magrin-Chagnolleau. *Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte*. PhD thesis, École Nationale Supérieure des Télécommunications, January 1997.

[3] Ivan Magrin-Chagnolleau and Geoffrey Durou. Application of time-frequency principal component analysis to speaker verification. *Digital Signal Processing*, 10(1-3), April 2000.

[4] A. Martin et al. The DET curve in assessment of detection task performance. In *Proceedings of EUROSPEECH 97*, volume 4, pages 1895–1898, September 1997. Rhodes, Greece.

[5] Geoffrey J. McLachlan and Kaye E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1988.

[6] Mark Przybocki and Alvin Martin. The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10(1-3), April 2000.

[7] The ELISA Consortium. The ELISA Systems for the NIST'99 Evaluation in Speaker Detection and Tracking. *Digital Signal Processing*, 10(1-3), April 2000.

[8] Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1–2):91–108, August 1995.

[9] Frank K. Soong and Biing-Hwang Juang Line Spectrum Pair (LSP) and Speech data compression. *ICASSP*, 1984.

[10] D. M. Titterington, A. F. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, 1985.