MODEL DEPENDENT SPECTRAL REPRESENTATIONS FOR SPEAKER RECOGNITION

G. Gravier † *C. Mokbel* ‡ *G. Chollet* †

†*ENST, Dpt. Signal, 46 rue Barrault, 75634 Paris Cedex 13, France*
‡*France Télécom, CNET - DIH/RCP, Lannion*
{*gravier,chollet*}*@sig.enst.fr mokbel@cnet.lannion.fr*

Abstract

We investigate the use of variable resolution spectral analysis for speaker recognition. The spectral resolution is simply determined by a unique parameter. A speaker can therefore be represented by this parameter and a stochastic model, which means that each speaker is represented in a different acoustic space. For speaker verification tasks, the likelihood ratio compared to a threshold should not depend on the representation space, so that likelihood ratios remain comparable. We experimented different spectral resolution with several classifiers but we had no improvement in the results and the classifiers turned out not to be very sensitive to the different feature sets.

1 INTRODUCTION

For both speech and speaker recognition, the signal is transformed into feature vectors, each vector representing the short term spectrum. Those vectors, belonging to an *acoustic space*, define a new representation of speech which is used for training models and for decoding. Many spectral representations have been studied (e.g. [1, 2]) but the most popular ones remain cepstral coefficients on linear (LPCC, LFCC) or Mel frequency (MFCC) scales. MFCC are mainly used for speech recognition since they offer a higher spectral resolution in low frequencies. Since speaker specific information relies more on higher frequencies, LPCC or LFCC are known to perform better than MFCC for speaker recognition. To our knowledge, very few studies have been carried out concerning frequency warping. Noda [3] showed that frequency-warped cepstra were robust to noise but considering a frequency warping for each phoneme.

We propose a study of various frequency warpings using variable resolution spectral analysis [4], for text-independent speaker verification. The spectral resolution being determined by one parameter, it is easy to include the latter in a speaker model. Therefore, each speaker is represented in a different acoustic space, and the speaker specific representation is determined so that the separability between the two test hypothesis (the speaker is who he/she claims to be vs. the speaker can be anyone) is maximum. These experiments are a first step toward a model dependent representation.

The paper is organized as follows: we first present variable resolution spectral analysis and the techniques that goes along with it. We then present in section 3 the motivations and the theoretical framework for the experiments. We describe the database and the experiment protocole in section 4 and we finally present the results and discuss them.

2 VARIABLE RESOLUTION SPECTRAL ANALYSIS

The principle of variable resolution spectral analysis is to apply a non-linear transformation to the frequency scale $\overline{\omega} = \Theta(\omega)$, controled by a parameter α, given in our case by:

$$\Theta(\omega) = \arctan\left|\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}\right| \qquad (1)$$

The frequency scale transformation is plotted for different values of alpha in figure 1. As can be seen, performing a classical spectral analysis on the $\overline{\omega}$ domain is equivalent to performing a variable resolution spectral analysis on the original domain. Positive values of α leads to a better low frequency resolution while negative values offers a better high frequency resolution.

There are two approaches to perform such an analysis. The first one is a parametric approach while the second one is based on filter-banks. Figure 2 illustrates variable resolution spectral analysis using both approach, for $\alpha = 0.0$ and $\alpha = 0.4$, filter bank outputs and spectrum enveloppe being represented in the original domain ω.

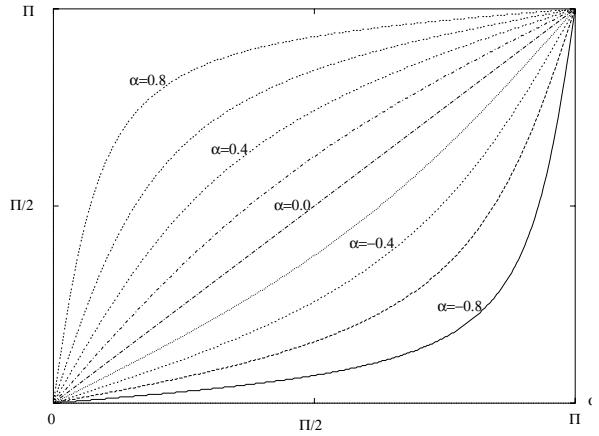

Figure 1: Spectral deformation as a function of α

2.1 Parametric approach

This approach is based on generalized linear prediction. For a pth-order analysis, the correlation sequence is computed at the output of a cascade of p first order dephasing filters rather than on the signal itself. The transfer function of the filters is given by $H(z) = \frac{z^{-1}-\alpha}{1-\alpha z^{-1}}$ and the cascade is preceded by a corrective filter $\mu(z) = \frac{1-\alpha^2}{(1-\alpha z^{-1})^2}$. The standard Levinson procedure is applied to the correlation sequence and yields prediction coefficients.

2.2 Non-parametric approach

As we know the frequency scale transformation given by eq. 1, it is also possible to determine the central frequencies of a filter bank, equally spaced on axis $\overline{\omega}$, and to use classical filter bank techniques based on the FFT of the signal.

3 TOWARD A SPEAKER DEPENDENT REPRESENTATION

If we denote X_α a feature vector obtained with the spectral resolution given by α, and X a feature vector in the reference acoustic space (i.e. $\alpha = 0$), we have the following relation:

$$p(X) = |det(J_\alpha)|.p(X_\alpha)$$

where $p(.)$ is the probability density function and J_α is the jacobian of the transformation $X \longmapsto X_\alpha$, assumed to be reversible. When using a normalisation cohort for speaker verification [5], the decision is taken by comparing the likelihood ratio of the two test hypothesis ($\frac{p_{H_0}(.)}{p_{H_1}(.)}$) to a threshold, where the hypothesis are:

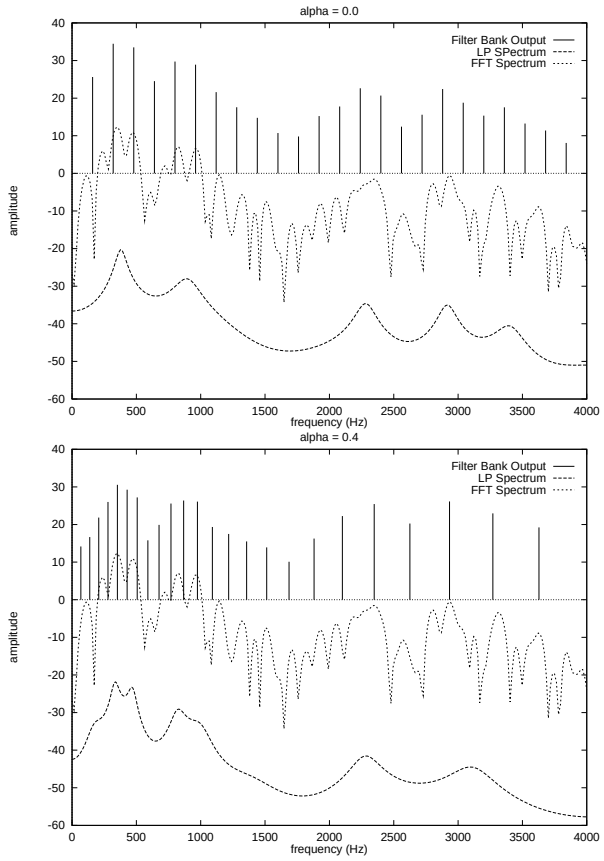

Figure 2: Examples of variable resolution analysis

- H_0: the speaker is the one he/she claims to be
- H_1: the speaker is not he/she claims to be

According to the above equation, and denoting λ_ω the model for H_1, we see that the likelihood ratio should not depend on α since:

$$\frac{p(X;\lambda_l)}{p(X;\lambda_\omega)} = \frac{p(X_\alpha;\lambda_l)}{p(X_\alpha;\lambda_\omega)} \qquad (2)$$

Though this relation is mathematically correct if the trqnsformation is reversible, we do not believe that it is exactly true in our case since the information captured in the signal under different spectral resolutions is not the same. Some speaker may be better represented in an acoustic space than in another one. But this relation shows that likelihood ratios in the different acoustic spaces are comparable. Each speaker model can therefore be defined in the most appropriate space for this speaker, provided the normalisation cohort model is defined on the same space.

4 APPLICATION

4.1 Databases

All our experiments are carried on the NIST 96 development database [6] for the one session training condition. This database is derived from SwitchBoard Corpus and contains 43 male and 45 female speakers. The training material consists of 2 minutes of speech recorded in a single session with a single handset. The test material consists of about 400 segments with three length conditions, namely 3, 10 and 30 seconds. Each test segment is matched against all same-sex speaker models and we have about 400 target-speaker trials and 17000 imposter trials, for each length of test segments.

An auxiliary database was used to train the speaker-independent normalization model. This database is also derived from SwitchBoard Corpus and contains about 80 speakers (half male, half female) speaking one minute each.

4.2 Feature vectors

We presented in section 2 two approaches for variable resolution spectral analysis. For generalized linear predictive analysis, cepstral coefficients can be determined from the prediction coefficients. For filter bank analysis, cepstral coefficients are calculated by mean of a discrete cosine transform. We used 16 cepstral coefficients computed on 32 ms windows after pre-emphasis and windowing, with a 10 ms frame shift. For the parametric approach we used a 16th order linear prediction. For the filter bank analysis, we used 24 triangular filters. First order derivatives of the cepstral coefficients and of the log-energy are appended to capture dynamic information in the signal, leading to a 33 dimensional feature vector. We also used cepstral mean substraction for channel compensation as well as liftering.

5 EXPERIMENTS AND RESULTS

5.1 Gaussian Mixture Models

We first studied the different acoustic spaces for $\alpha \in \{-0.4, -0.2, 0.0, 0.2, 0.4\}$ using Gaussian Mixture Models [7] and the parametric approach for feature extraction (LP-GMM). Speaker models have 32 components while the normalisation model has 128 components. The equal error rates are listed in table 1 and results obtained with Mel frequency cepstral coefficients are given for comparison.

	Test segment duration		
	3 sec.	10 sec.	30 sec.
$MFCC$	18.71	15.76	12.52
$\alpha = -0.4$	20.0	14.54	13.04
$\alpha = -0.2$	18.80	14.27	13.04
$\alpha = 0.0$	19.47	13.76	13.70
$\alpha = 0.2$	19.69	14.77	13.08
$\alpha = 0.4$	20.70	17.22	15.03

Table 1: Equal Error rates (in %) for LP-GMM

It can be noticed that there is no significant difference between the spectral representations. However with excessive spectral deformation, the error rate increases. For the 3 second test segments, we measured EERs of 29.12% and 22.98% for $\alpha = -0.8$ and 0.8 respectively. It seems that the information extracted from the signal and represented by the gaussian mixture is the same as long as the spectral transformation is not excessive.

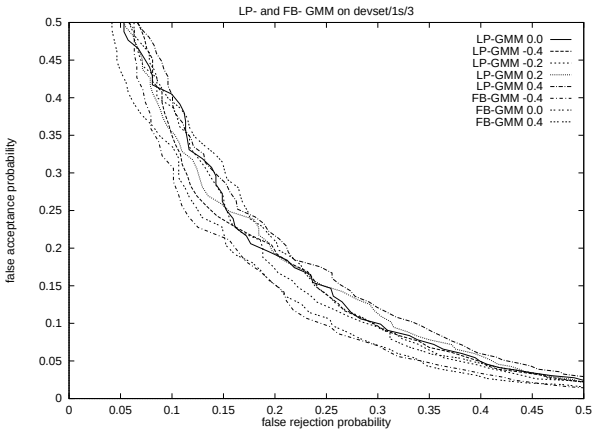

Figure 3: LP- and FB-GMM ROC (3 sec.)

We also tested the non-parametric approach for feature extraction (FB-GMM). Results for the 3 second tests are given for three values of α in table 2. Filterbank derived features perform globally better than LP-derived ones but still there is no significant differences in the results with various resolutions. ROC curves are given for the 3 second test segments in figure 3. However, it can be noted that, in both case, results seems to deteriorate for $\alpha = 0.4$ while they remain constant for $\alpha = -0.4$. This result outlines the importance of high frequencies for speaker recognition and suggests that a slight difference exists between the feature sets.

α	-0.4	0.0	0.4
FB-GMM	17.70	17.94	20.46

Table 2: Equal Error rates (in %) for FB-GMM

5.2 Second order measures

Both LP- and FB-GMM experiments showed that there is no significant difference between the spectral representations we studied. We see two possible reasons for this. Either the information captured in the signal remains the same or the model extracts the same information from the feature vectors. To answer that question, we experimented second order statistical measures [8] on the 3 second test segments with LP-derived features (LP-SOSM). Actually the model can be seen as a single state, mono-gaussian HMM with a full covariance matrix. Results are listed in table 3. Again, except for $\alpha = 0.4$, the EERs are not significantly different.

α	-0.4	-0.2	0.0	0.2	0.4
SOSM	17.66	16.98	17.13	19.37	22.77

Table 3: Equal Error rates (in %) for LP-SOSM

5.3 Speaker specific representation

Finally, we tested the speaker specific spectral representation approach with LP-derived features and GMMs. For each speaker, we choose the parameter α that maximises the likelihood ratio on the training corpus. In other words, we select the representation for which the two test hypothesis are the most distinct. Results are poor since the EER is 16.93% for the 10 second tests and 20.90% for the 3 second ones. One of the reason we supposed was that equation 2 did not apply in our case and that the problem was due to the use of a global threshold. A-posteriori EERs determined separatly for each value of α showed that the poor performances were mainly due to the criterion used to select α. For $\alpha = -0.2$ the EER is 14.63% but it is 21.31% for $\alpha = -0.4$, and the mean value is 18.66%.

6 Conclusion

In this paper, we investigated the use of variable resolution spectral analysis in speaker recognition and tried to define a speaker specific representation using this technique. The experiments on text-free speaker verification showed that a slight difference exists between the spectral representations but that it does not increase significantly the performance of the system. However, this study is a first step toward a model dependent representation of speech and outlines that the selection of a suitable representation for a speaker is not obvious.

References

[1] Chafic Mokbel. *Reconnaissance de la parole dans le bruit: Bruitage / Débruitage.* PhD thesis, Ecole Nationale Superieure des Télécommunications, 1992.

[2] M. Homayounpour and G. Chollet. A comparison of some relevant parametric representations for speaker verification. In *ESCA Workshop on Speaker Recognition, Identification and Verification*, pages 185–188, 1994.

[3] Hideki Noda. Frequency-warped spectral distance measures for speaker verification in noise. In *ICASSP*, volume 1, pages 576–579, 1988.

[4] Christian Chouzenoux. *Analyse spectrale à résolution variable. Application au signal de parole.* PhD thesis, Ecole Nationale Superieure des Télécommunications, 1982.

[5] Aaron E. Rosenberg et al. The use of cohort normalized scores for speaker verification. In *International Conference on Speech and Language Processing*, pages 599–602, 1992.

[6] *NIST Speaker Recognition Workshop*, 1996.

[7] Douglas A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *ESCA Workshop on Speaker Recognition, Identification and Verification*, pages 27–30, 1994.

[8] F. Bimbot et al. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, 17:177–192, 1995.