IDIAP RESEARCH REPORT

An Investigation of Spectral Subband Centroids for Speaker Authentication

Norman Poh Hoon Thian [a]

Conrad Sanderson [a] Samy Bengio [a]

IDIAP–RR 03-62

December 2003

PUBLISHED IN

International Conference on Biometric Authentication (ICBA 2004), Hong Kong, pages 631-639, July 15-17, 2004.

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP, CP 592, 1920 Martigny, Switzerland

IDIAP Research Report 03-62

An Investigation of Spectral Subband Centroids for Speaker Authentication

Norman Poh Hoon Thian Conrad Sanderson Samy Bengio

December 2003

published in
International Conference on Biometric Authentication (ICBA 2004),
Hong Kong, pages 631-639, July 15-17, 2004.

Abstract. Most conventional features used in speaker authentication are based on estimation of spectral envelopes in one way or another, in the form of cepstrums, e.g., Mel-scale Filterbank Cepstrum Coefficients (MFCCs), Linear-scale Filterbank Cepstrum Coefficients (LFCCs) and Relative Spectral Perceptual Linear Prediction (RASTA-PLP). In this study, Spectral Subband Centroids (SSCs) are examined. These features are the centroid frequency in each subband. They have properties similar to the formant frequency but are limited to a given subband. Preliminary empirical findings, on a subset of the XM2VTS database, using Analysis of Variance and Linear Discriminant Analysis suggest that, firstly, a certain number of centroids (up to about 16) are necessary to cover enough information about the speaker's identity; and secondly, that SSCs could provide complementary information to the conventional MFCCs. Theoretical findings suggest that mean-subtracted SSCs are more robust to additive noise. Further empirical experiments carried out on the more realistic NIST2001 database using SSCs, MFCCs (respectively LFCCs) and their combinations by concatenation suggest that SSCs are indeed robust and complementary features to conventional MFCC (respectively LFCCs) features often used in speaker authentication.

1 Introduction

Speech recognition is the task of determining the linguistic contents of a speech signal, while speaker authentication is the task of verifying whether a person really is who he or she claims to be. Even though both tasks are very different, the front-end processing of speech signals is often common. Although there exists some research efforts in designing new and effective speech features for speaker authentication [10] (i.e., Line Spectrum Pairs, Time-Frequency Principal Component and Discriminant Components of the Spectrum), Mel-scale Frequency Cepstral Coefficients (MFCCs) features, which are commonly used in speech recognition, remain the state-of-the-art features, as long as speaker authentication is concerned. Empirical studies in [14] showed that Linear-scale Frequency Cepstral Coefficients [13] (LFCCs) achieve comparable performance to that of MFCCs [14]. According to the same study, Perceptual Linear Prediction (PLP) cepstral coefficients, which are widely used in speech recognition, did not perform significantly better than MFCCs. Furthermore, in the same experiment setting, the performance of PLP with RASTA-preprocessing [7] (RASTA-PLP) was slightly worse than PLP alone. Hence, features that work better in speech recognition *may not* always work better in speaker authentication.

The aim of this study is double-fold: to provide complementary features that describe information not captured by the conventional state-of-the-art MFCC features for speaker authentication tasks; and to examine how these features perform alone, as compared to MFCC features. One such information is the spectral information. In [1, Sec. 3.3], frequency and amplitude information are extracted from "spectral lines" [4]. Spectral lines are extracted from the spectrogram of a signal by using thinning and skeletonisation algorithms that are often used in image-processing. Low frequency spectral lines in this case actually correspond to the fundamental frequency or pitch. The pair (frequency, amplitude) hence represents a point on this 2D space. With quantisation on frequency and amplitude, this frequency/amplitude encoded data is classified using a feed-forward network and is shown to achieve lower generalisation error as compared to the encoding scheme which uses fixed frequency intervals with their corresponding amplitude values. The study suggests that frequency information, when encoded properly, can increase the robustness of a *speech recognition* system.

Contrary to the first approach, in the context of *speaker authentication*, Sönmez *et al* directly estimated the (long-term) pitch information using parametric models called log-normal tied mixture model [17]. A follow-up work [16] used the (local variation of) pitch dynamics which contain speaker's intonation (speaking style). In both works, the resultant pitch system is combined with the cepstral feature-based system by summation of (log-)likelihood scores over the same utterance. They all show improvement over the baseline system.

In the context of *speech recognition*, frequency information is represented in the form of Spectral Subband Centroids (SSCs) [12], which represent the centroid frequency in each subband. In conventional MFCC features, the power spectrum in a given subband is often smoothed out, so that only the (weighted) amplitude of the power spectrum is kept. Therefore, SSCs provide different information to conventional MFCCs. It has been demonstrated [12] that SSCs, when used in conjunction with MFCCs, result in better speech recognition accuracy than that of the baseline MFCCs; when used alone, SSCs achieve performance that is comparable (but with slight degradation) to that of MFCCs.

Would frequency information enhance the performance of a speaker authentication system? According to [16, 17], the answer is yes. How should this information be incorporated into an existing system based on MFCC features? In this work, SSCs are used as a preliminary study because they can be incorporated at the frame-level (and of course at classifier-score level) while this is not possible in [16, 17]. Furthermore, in these works, spectral information other than pitch (e.g. higher frequency band) is not used at all. Secondly, SSCs have not been applied to speaker authentication, which in this case, constitute an interesting research question.

The rest of this paper is organised as follows: Section 2 explains briefly the experiment setting. Analysis of SSCs is discussed in Section 3. This is followed by empirical results in Section 4 and conclusions in Section 5.

2 Experiment Setup

As a preliminary study, subsets of XM2VTS and NIST2001 are used. XM2VTS is used here to study the features under clean conditions and with only limited vocabulary (i.e., digits only). NIST2001 is used to evaluate how well these features perform on telephone data with and without additive environmental noise, on speaker authentication tasks.

In the case of XM2VTS [9], 6 male and 6 female speakers were chosen randomly. For each speaker, 8 recordings are available from 4 sessions and each session has 2 utterances. Sessions were recorded at one-month intervals. As for the NIST2001 database [11], which comes from the Switchboard-2 Phase 3 Corpus collected by the Linguistic Data Consortium, only the female subset (which is known to be slightly more difficult than the male subset) is used for evaluation. In the original database two different handsets were used (i.e., carbon and electret). However, only data from electret handsets are used (5 speakers who used the carbon handsets were removed) so that any variation of performance, if any, will not be attributed to this factor. This database was separated into three subsets: a training set for the world model, a development set and an evaluation set. The female world model was trained on 218 speakers for a total of 3 hours of speech. For both development and evaluation (female) clients, there was about 2 minutes of telephone speech used to train the models and each test access was less than 1 minute long. The development population consisted of 45 females while there were 506 females in the evaluation set. The total number of accesses for the development population was 2694 and 32029 for the evaluation population with a proportion of 10% of true accesses. Four types of noise (`white`, `oproom` (for operational room), `factory` and `lynx`), taken from the NOISEX-92 database [19], were used to contaminate the NIST2001 dataset.

The classifier used in this paper is based on Gaussian Mixture Models (GMMs). It models the statistical distribution of training feature vectors for each client. Given a claim for client C's identity and a set of (test) feature vectors $X = \{\vec{x}_i\}_{i=1}^{N_V}$ supporting the claim, the average log likelihood (over N_V feature vectors) of the claimant being the true claimant is found with:

$$\mathcal{L}(X|\lambda_C) = \frac{1}{N_V} \sum\nolimits_{i=1}^{N_V} \log p(\vec{x}_i|\lambda_C), \tag{1}$$

where λ_C is a set of GMM parameters associated with client C. Given the average log likelihood of the claimant being an impostor, the opinion on the claim is found using average Log Likelihood Ratio (LLR), as follows:

$$\mathrm{LLR}(X) = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \tag{2}$$

In its general form, a GMM model with parameter λ can be described by:

$$\begin{aligned} p(\vec{x}|\lambda) &= \sum\nolimits_{j=1}^{N_G} w_j \ \mathcal{N}(\vec{x}; \vec{\mu}_j, \mathbf{\Sigma}_j), && (3) \\ \lambda &= \{w_j, \vec{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{N_G}. && (4) \end{aligned}$$

where $\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma})$ is a D-dimensional Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$, N_G is the number of Gaussians and w_j is the weight for Gaussian j (with constraints $\sum_{j=1}^{N_G} w_j = 1$ and $\forall\ j : w_j \geq 0$). A common impostor GMM (also called a world or universal background model [15]) is used to model the statistics of the mentioned 218 female speakers. It is trained using the Expectation-Maximization (EM) algorithm [2]. This world model is then adapted to each client's speech features using Maximum *a Posteriori* (MAP) estimation [15]. The world model was used to evaluate the hypothesis of an impostor's access while a client-adapted model was used to evaluate the hypothesis of a client's access. To make a decision, average LLR (Eqn. (2)) is compared to a threshold chosen on a development data.

The commonly used Half Total Error Rate (HTER) is used as evaluation criterion. It is defined as (FAR + FRR)/2, where FAR is False Acceptance Rate and FRR is False Rejection Rate. Here, we assume that the costs of false acceptance and false rejection are equal and that the prior (class) distribution of clients and impostors are equal as well. The HTER is calculated based on a threshold which itself is estimated *from a development set.* This threshold is estimated such that $|\text{FAR}(\theta) - \text{FRR}(\theta)|$ is minimised with respect to θ. It is then used to make decisions on an evaluation set. Hence, the HTER is *unbiased* with respect to the evaluation set since its associated threshold is estimated *a priori* on the development set. We call the resultant measure an *a priori* HTER and is used whenever an evaluation set is used. The smaller HTER is, the better the classification result.

3 SSCs: An Investigation

This section contains several studies on SSCs: Section 3.1 introduces the basic ideas of SSCs. Section 3.2 discusses the use of analysis of variance on SSCs to determine the number of centroids[1]. This is confirmed by an empirical study reported in Section 3.3. Section 3.4 justifies why mean-subtraction can also be applied to SSCs. A preliminary series of experiments is reported in Section 3.5 to demonstrate empirically that mean-subtraction and temporal information (i.e., two techniques which are commonly used in MFCC features) can also be applied to SSCs to improve the performance.

3.1 SSC Features

Let the frequency band $[0, F_s/2]$ be divided into M subbands, where F_s is the sampling frequency. For the m-th subband, let its lower and higher edges be l_m and h_m, respectively. Furthermore, let the filter shape be $w_m(f)$ and $P^\gamma(f)$ be the power spectrum at location f raised to the power of γ. The m-th subband centroid, according to [12], is defined as:

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^\gamma(f) df}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f) df}. \tag{5}$$

Note that the term $w_m(f)P^\gamma(f)$ can be viewed as a bias which influences where the centroid should be. A peak in this term leads to a higher weight in the corresponding f. Typically, $w_m(f)$ takes on the shape of either a square window (ones over the m-th subband and zeros everywhere else) or a triangular window (which gives a maximum response around its center and decreases towards both of it edges). In the case of MFCCs, w_m is a triangular window. This same window is used here. The use of γ parameter in this function is rather a design parameter and is not motivated by any psychological aspect of hearing. The γ parameter has been used elsewhere in the literature [3] as part of feature extraction (which is called a two-dimensional root spectrum) for speech recognition. According to that study, γ is a design parameter which can be optimised on a given data set and task at hand. Hence, the introduction of γ is only for practical reasons from engineering point of view. In this report, γ is set to 1.

Figure 1 shows a conventional spectrogram overlaid with the SSC features with 5 equally-spaced bands, calculated using square windows. This is done so to verify what exactly SSCs represent. This utterance contains three digits: zero, one and two. It can be observed that, firstly, when there is no speech, SSCs in a given frequency subband tend to be the center of the band. On the other hand, with the presence of speech, SSCs show some regular trends: the trajectory of SSCs in a given subband actually locates the peaks of the power spectrum limited in that given subband. This coincides with the idea of spectral lines [4] discussed earlier. However, in this context, the representation is limited to one value per subband. Secondly, the medium to long-term time-trajectory of SSCs can be an interesting feature set as well, as demonstrated in [16]; this is not the subject of study here but will be examined in the future. Thirdly, if there are not enough centroids, then SSCs will not cover enough

[1] Additional experiments based on Linear Discriminant Analysis for testing class-separability and complementary based on SSCs and MFCCs, and others based on F-ratio tests of these two features can be found in the Appendix. The results of these findings suggest that class labels (speaker's identities) not separable in one feature space are probably separable in another feature space; and that the feature space induced by MFCCs is more separable than that induced by SSCs, thus predicting that the performance due to MFCCs under matched conditions is probably better than that due to SSCs.

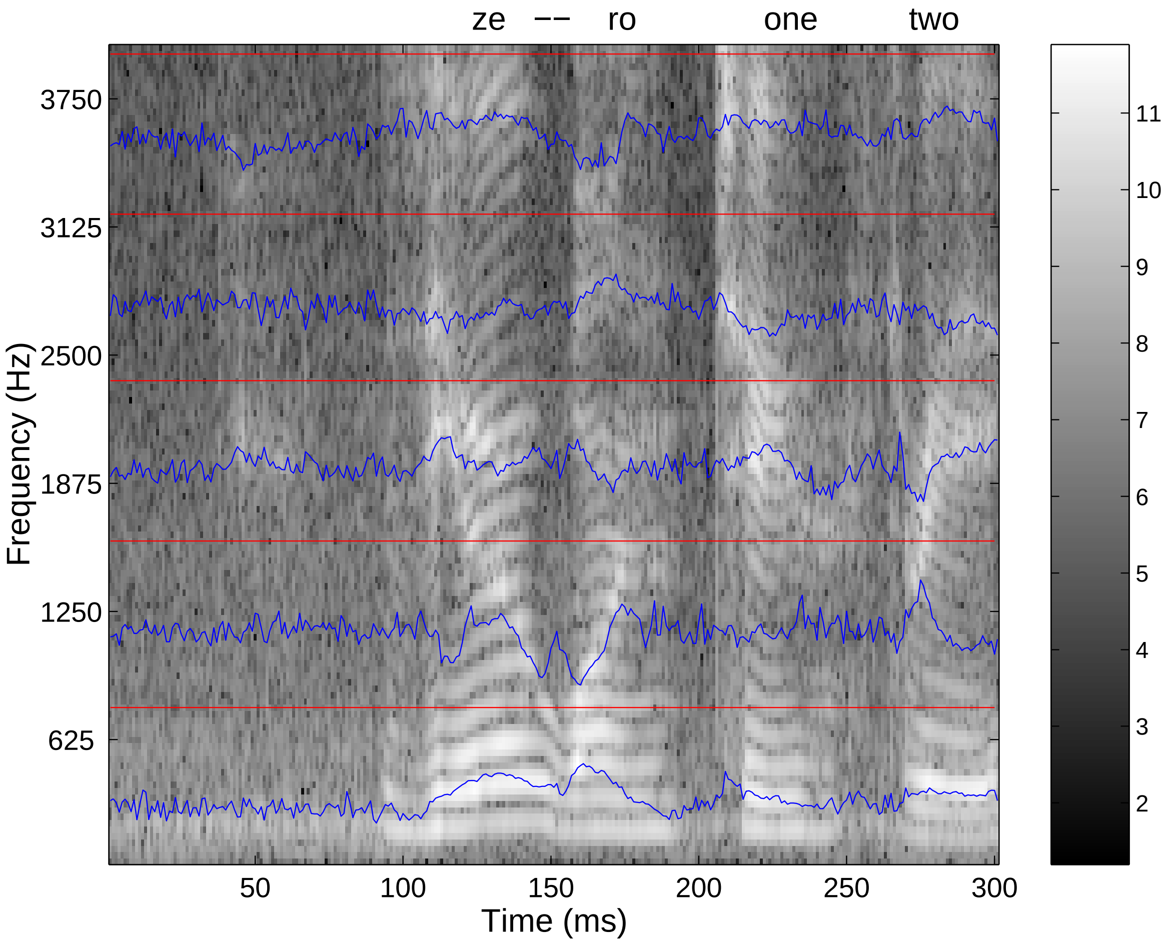

Figure 1: SSC features across time

information. On the other hand, if there are too many centroids, additional centroids will only add to the unnecessary dimensionality of the data, without adding any more information. Therefore, it is important to measure the amount of information covered using techniques discussed in the next subsection.

3.2 ANOVA Analysis of SSCs and MFCCs

Linear Discriminant Analysis (LDA), under the Analysis of Variance (ANOVA) framework has already been used elsewhere in the literature [8], in the context of speech recognition. LDA assumes that the sources captured by the speech features have a unimodal multi-variate Gaussian distribution. These sources could be useful or harmful. In the context of *text-independent* speaker authentication, useful variance of sources are speaker's intonation, habitual speaking rate and physical articulatory features, while harmful sources of variance are linguistic content, context variability (e.g. free-conversation, telephone interview), environmental (hence assumed additive) noise and channel (assumed additive and/or convolutional) noise (e.g. handset difference, telephone-and microphone-induced room acoustics, reverberation, etc) [6, Section 4.1]. However in this context, we do not use LDA to design discriminative features but firstly, to approximate the number of centroids that is needed to obtain optimal features; and secondly, to analyse and visualise the interaction between the within- and between-class variance.

Table 1: LDA analysis

No. of coefficients/ centroids	No. of LDA coefs capturing 99% of var.	
	SSC	MFCC
4	4	4
8	6	7
12	8	9
16	9	9
24	9	9
30	9	–

6 male and 6 female speakers, each with 8 utterances of digits, taken from XM2VTS are used here. Each feature frame is labeled with the speaker's identity. The first analysis consists of finding the number of optimal centroids in SSCs. This is done as follows:

1. Compute SSC features using M bands
2. Label each feature frame according to the speaker's identity
3. Calculate the between-class variance Σ_b and the within-class variance Σ_w, both with dimension $M \times M$.
4. Calculate the eigen-values λ by solving $eig(\Sigma_w^{-1}\Sigma_b)$.
5. Find the number of eigen-values d needed to accumulate 99% of variance, i.e., $\arg\min_d \Big\{ d : d \in \mathbb{N}, 1 \leq d \leq M, \sum_{i=1}^{d} \lambda_i / \sum_{i=1}^{M} \lambda_i \geq 0.99 \Big\}$.

This procedure is repeated for different values of M in increasing order. We also perform the same analysis for MFCCs with different numbers of coefficients. Here, the number of filter-banks is fixed at 24. The results are shown in Table 1. Since there are 12 speakers, we note that d can be at most 11. As can be seen, for the case of SSCs, at first, 4 centroids are shown to be mutually independent from each other. That is why in the LDA space, all 4 orthogonal vectors are needed to describe (99% of variance of) discriminant directions that best separate the speakers. However, as more centroids are used, less and less orthogonal vectors are needed as compared to the number of centroids used. We note that the number of orthogonal vectors (or the LDA eigen-values that capture 99% of variance) stays at about 9 (which is less than 11) even though the number of centroids increases beyond 16. This shows that additional information due to adding the number of centroids beyond 16 is not likely to contain more information that what has already been covered. Hence, excessive number of centroids will add unnecessarily to computational cost on the part of the classifier. As for the MFCCs, the same explanation as in SSCs applies. Note that in the last row, it is not possible to derive 30 coefficients from 24 filter-banks. The maximum number of coefficients possible is 24 in this case. Also, according to the LDA analysis, only 9 orthogonal vectors (even though 11 is the maximum) are used to describe discriminant directions for speaker authentication even though the number of coefficients is increased from 12 to 24. This would probably explain why 12 to 24 coefficients are commonly used in speech and speaker recognition. We conjecture that the relevant number of coefficients are dependent on a given task and a given dataset (which can be tuned by empirical procedures such as cross-validation).

3.3 The Number of Centroids: An Empirical Study

To further verify the number of centroids to use, several experiments are conducted by varying the number of centroids[2]. Mean-subtraction is applied here to compensate for channel noise. In fact,

[2] The silence-speech segmentation is not optimal in this experiment. Furthermore, delta features are not added, and 0-4000 Hz frequency is used instead of 300-3400 Hz. Hence, the result is sub-optimal. Here, the purpose is to give an

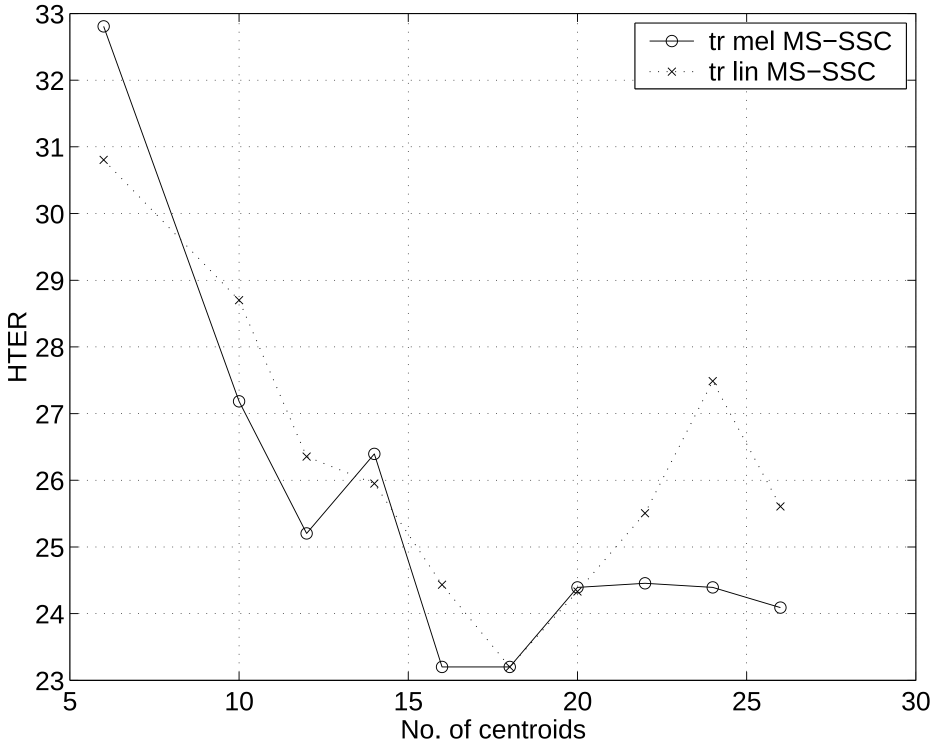

Figure 2: *A posteriori* HTERs (in %) of SSCs (with mean-subtraction) using different number of centroids on the female development subset of NIST2001 database

without mean-subtraction, the results would have been unacceptably high (a justification of mean-subtraction is given in Section 3.4). Two types of scales are used: Mel-scale and linear-scale which are labeled as "tr mel" and "tr lin" in Figure 2 respectively. The filter-bank windows are triangular windows centered on the critical band. Throughout the experiments, 128 Gaussian components are used (see Section 2). This number is chosen by cross-validation based on LFCCs-induced GMMs. The HTER curves for the two window types exhibit a general trend: as the number of centroids increases, HTER reduces and reaches a minimum around 16-18 centroids before increasing again. This is somewhat in accordance with the LDA analysis done in Section 3.2.

3.4 Resistance to Noise with Mean Subtraction

Since SSCs are derived from the power spectrum, they are also affected by channel and additive noise. In this section, we will show how noise is compensated via SSCs mean subtraction.

Let $S(n, \omega)$ be the result of applying Fourier Transform on the n-th windowed signal, where ω is the discretised frequency (in radians). Furthermore, let the power spectrum of $S(n, \omega)$ be $\tilde{S}_n(f) = |S(n, \omega)|^2$, where $f = \omega/2\pi$. $\tilde{S}_n(f)$ is often called a spectrogram. This term will be referred to as $\tilde{S}_n$ here. The decomposition of the signal in the Fourier domain will be:

$$\tilde{S}_n = \tilde{S}_{o,n} + \tilde{V}_n, \tag{6}$$

where $\tilde{S}_{o,n}$ is the power spectrum of the clean speech signal and $\tilde{V}_n$ is that of the additive noise

idea of how the number of centroids can influence the overall *relative* performance.

component. We can rewrite Eqn. (5) as:

$$C \;=\; \frac{\int f\tilde{S}_n(f)df}{\int \tilde{S}_n(f)df} = \frac{\int f\tilde{S}_{o,n}(f)df + \int \tilde{V}_n(f)df}{\int \tilde{S}_n(f)df}, \tag{7}$$

where, C has the same meaning as in Eqn. (5); the subscript m and filterbank edges l_m and h_m are omitted. The expected value of C with respect to time (n) can be calculated as:

$$\mathrm{E}_n\{C\} = \mathrm{E}_n\left\{\frac{\int f\tilde{S}_{o,n}(f)df}{\int \tilde{S}_n(f)df}\right\} + \mathrm{E}_n\left\{\frac{\int f\tilde{V}_n(f)df}{\int \tilde{S}_n(f)df}\right\}.$$

Then $C - \mathrm{E}_n\{C\}$ would be:

$$\underbrace{\frac{\int f\tilde{S}_o(f)df}{\int \tilde{S}_n(f)df} - \mathrm{E}_n\left\{\frac{\int f\tilde{S}_{o,n}(f)df}{\int \tilde{S}_n(f)df}\right\}} + \underbrace{\frac{\int f\tilde{V}_n(f)df}{\int \tilde{S}_n(f)df} - \mathrm{E}_n\left\{\frac{\int f\tilde{V}_n(f)df}{\int \tilde{S}_n(f)df}\right\}}. \tag{8}$$

Assuming that the additive component of the noise is stationary, i.e., $\tilde{V}_n(f) = \tilde{V}(f)$, the second underbraced term can be simplified to:

$$\int f\tilde{V}(f)df\left(\frac{1}{\int \tilde{S}_n(f)df} - \mathrm{E}_n\left\{\frac{1}{\int \tilde{S}_n(f)df}\right\}\right),$$

$$\int f\tilde{V}(f)df\left(\frac{1}{\int \tilde{S}_{o,n}(f)df + \mathrm{K}} - \frac{1}{\int \mathrm{E}_n\{\tilde{S}_{o,n}(f)\}df + \mathrm{K}}\right),$$

where, $\mathrm{K} = \int \tilde{V}(f)df$ is introduced into the term.

When $\mathrm{K} \gg \int \tilde{S}_{o,n}(f)df$ (i.e., the SNR is very small), the second underbraced term will approach zero. This will leave the first underbraced term only in Eqn. (8). The denominator

$$\int \tilde{S}_n(f)df = \int \tilde{S}o, n(f) + \int \tilde{V}_n(f)df$$

is unfortunately still influenced by the additive noise. Therefore, mean-subtracted SSCs can only compensate for additive noise but cannot remove it completely.

On the other hand, since a large portion of noise has already been canceled due to the second underbraced term of Eqn. (8), the net effect is that $C-\mathrm{E}_n\{C\}$ is more robust than C. In fact, empirical results in Section 3.5 confirm that without mean-subtraction, SSCs' performance is unacceptable in practice. Whether SSCs can compensate convolutional noise or not is an open research issue not dealt here.

3.5 Investigation of Deltas and Mean-Subtraction on SSCs: An Empirical Study

Besides mean-subtraction, delta information is also known to carry useful information [18]. Several experiments are carried out and the results are shown in Table 2. MFCCs with Cepstral Mean Subtraction (CMS) perform better than MFCCs without CMS. This shows that the NIST2001 database contains some unknown telephone channel noise. The performance of the baseline SSCs with 16 centroids is 34.4% of HTER. By adding delta information, the system achieves a HTER close to 30%. Similar trend is observed with systems using mean-subtracted version of SSCs. Furthermore, the latter two systems are better than the former two systems. Hence, we conclude that mean-subtraction is also useful for SSCs.

Table 2: *A posteriori* HTERs (in %) of SSCs and MFCCs with different post-processing evaluated on the female development subset of NIST2001 database

Features	HTER (%)
MFCC(24/12), delta	27.568
MFCC(24/12), CMS, delta	16.747
SSC(16)	34.405
SSC(16), delta,	29.996
SSC(16), MS	19.600
SSC(16), MS, delta,	17.192

4 Empirical Results in Mismatched Conditions

Preliminary studies in Section 3 show that the following configuration of SSCs is probably optimal for speaker authentication task: 16 centroids, sampled using triangular windows and spaced linearly on the Mel-scale, with delta information and mean-subtraction. This configuration was used on the female evaluation subset (contrary to the development subset used in Section 3). Furthermore, only bands in the 300-3400 Hz frequency range are used. The log of delta energy is also used. The absolute log energy is not used as a form of energy normalisation. There are two goals: to investigate how resistant SSCs are to mismatched noisy conditions; and to see if concatenation of SSCs with conventional features will help in authentication. Two conventional features are used here: LFCCs and MFCCs. The LFCCs are extracted using 24 filterbanks with 16 cepstrum coefficients. MFCCs are extracted using 24 filterbanks with 12 cepstrum coefficients[3]. Several noise types are artificially added to the database at the following Signal-to-Noise Ratios (SNRs): 18, 12, 6 and 0 decibels. Two sets of experiments are conducted: in the first set, MFCCs, SSCs and their combined features are trained in clean conditions and tested in different noisy conditions. Hence the combined MFCC-SSC features have $12 + 16 = 28$ dimensions. With delta information, which also has 28 dimensions and log energy, the resultant features have 57 $(28 \times 2 + 1)$ dimensions. Using the same configuration, the second set of experiments used LFCCs instead. The resultant LFCC-SSC combined features have 65 $((16 + 16) \times 2 + 1)$ dimensions. GMMs with 128 Gaussians are used as back-end classifiers for all experiments. The number of Gaussians was found by cross-validation based on the LFCCs features. The results are shown in Figures 3 and 4 for these two sets of experiments. For both sets of experiments, it can be observed that MFCCs (respectively LFCCs) perform better than SSCs under clean conditions but not as good as SSCs under noisy conditions. When MFCCs (respectively LFCCs) are combined with SSCs, the resultant feature sets perform better than any of the features when used alone, in both clean and noisy conditions. Hence, SSCs are potentially useful as complementary features for speaker authentication.

5 Conclusions

Spectral Subband Centroids (SSCs) are a recent set of features that exploit the dominant frequency in each subband. The use of SSCs in recent literature has shown some successes in speech recognition. In this study, the potential use of SSCs in *text-independent speaker authentication* task was studied using analysis of variance and Linear Discriminant Analysis on a small digits datasets taken from the XM2VTS database. Experiments done on the female development subset of the NIST2001 Switch-

[3] This configuration gives 16.747% *a posteriori* HTER on the NIST2001 development set. Another commonly used configuration has the following parameters: 24 filterbanks with cepstrums $C_1, \ldots, C_{23}$, hence throwing out C_0. Although the latter configuration is more robust to noise, in this experiment, its performance is 17.131% *a posteriori* HTER in clean conditions.

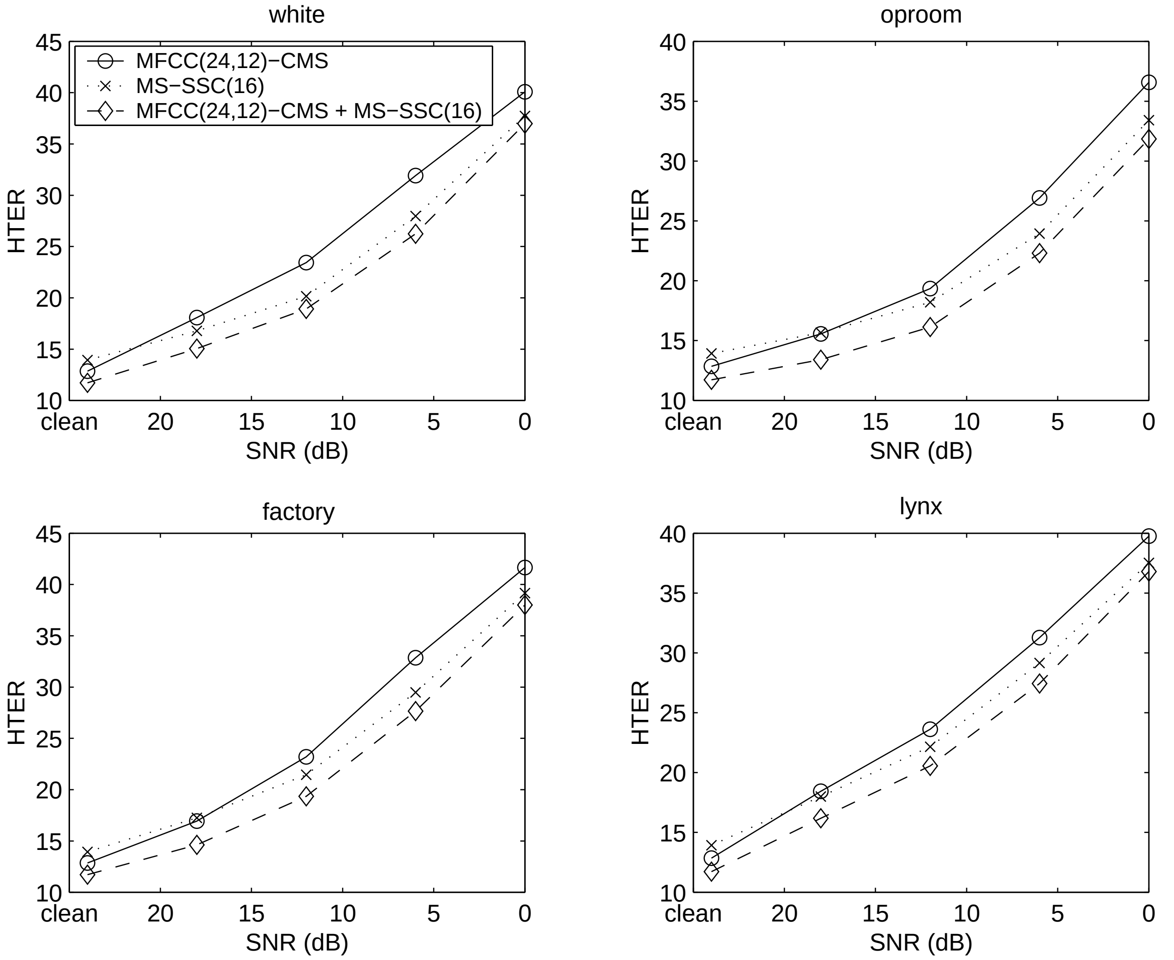

Figure 3: *A priori* HTERs (in %) of SSCs, MFCCs and MFCC+SSC feature sets on the female evaluation subset of NIST2001 database, under mismatched conditions, using threshold estimated from clean development data

Board database were used at the same time to validate the findings using ANOVA and LDA. Hence, based on the Half Total Error Rate (HTER), we can conclude that:

1. Mean-subtracted SSCs (MS-SSCs) with triangular windows on Mel-scale filterbank with 16 subbands give optimal results for speaker authentication;

2. MS-SSCs are more resistant to noise than the original SSCs;

3. MS-SSCs' first temporal derivative features are an important secondary set of features;

4. MS-SSCs are slightly inferior to conventional MFCCs with cepstral mean subtraction in clean conditions

Moreover, experiments done on the female evaluation subset of NIST2001 database, using the above SSCs configuration, showed that SSCs perform somewhat better than MFCCs in noisy conditions; and that combining SSCs with MFCCs (and respectively LFCCs) improves the accuracy of the system in both clean and noisy conditions compared to using any of the feature sets alone.

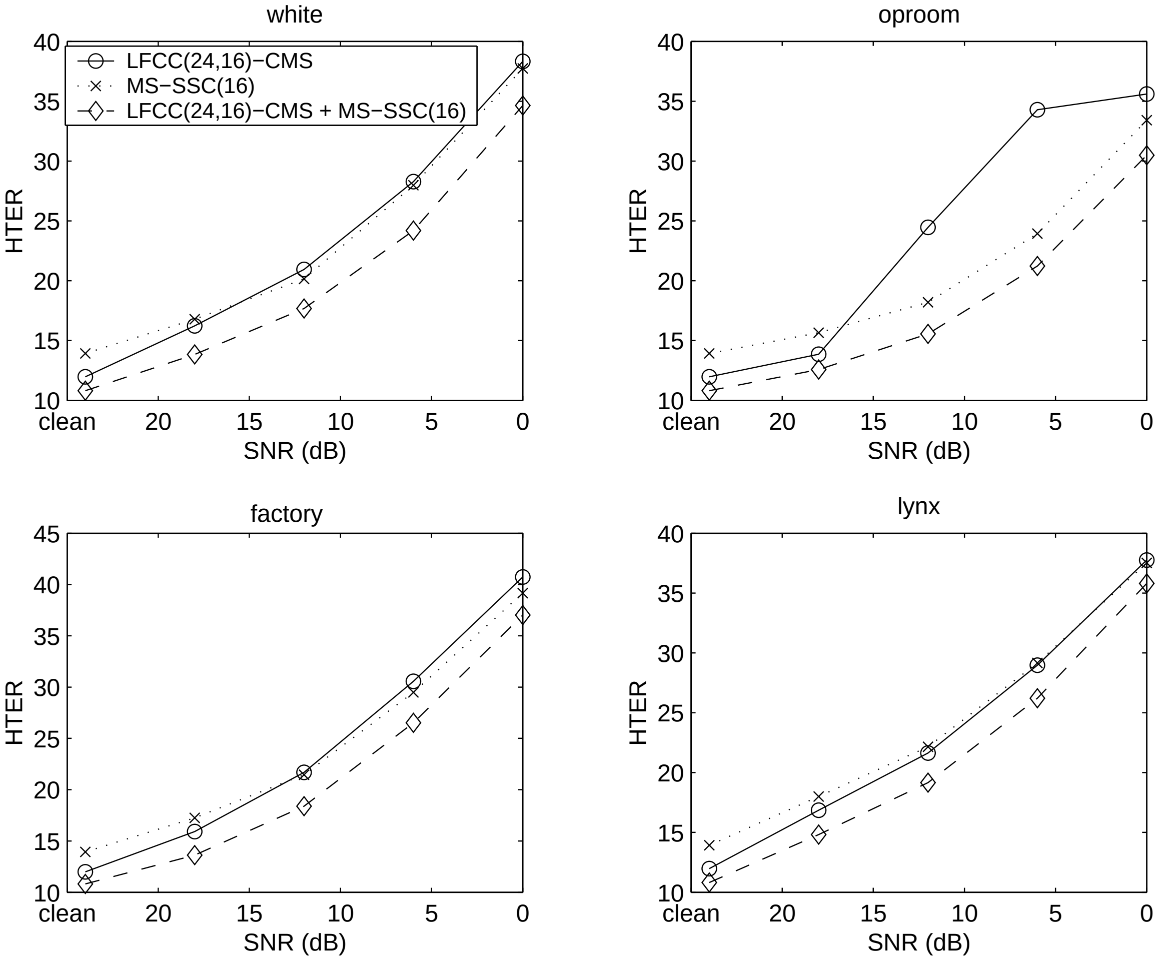

Figure 4: *A priori* HTERs (in (%) of SSCs, LFCCs and LFCC+SSC feature sets on the female evaluation subset of NIST2001 database, under mismatched conditions, using threshold estimated from clean development data

Hence, dominant frequencies represented by SSCs indeed contain speaker discriminative information, different from what MFCCs (respectively LFCCs) provide. In short, SSCs feature are a complementary and yet robust set of features for speaker authentication.

Acknowledgement

The authors thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". The authors also thank Hynek Hermansky for his constructive comments and suggestions.

APPENDICES

The following two analyses are based on a subset of XM2VTS discussed in Section 2. The goals of these additional studies are two-fold: to visualise the class-separability of MFCCs and SSCs, and to test

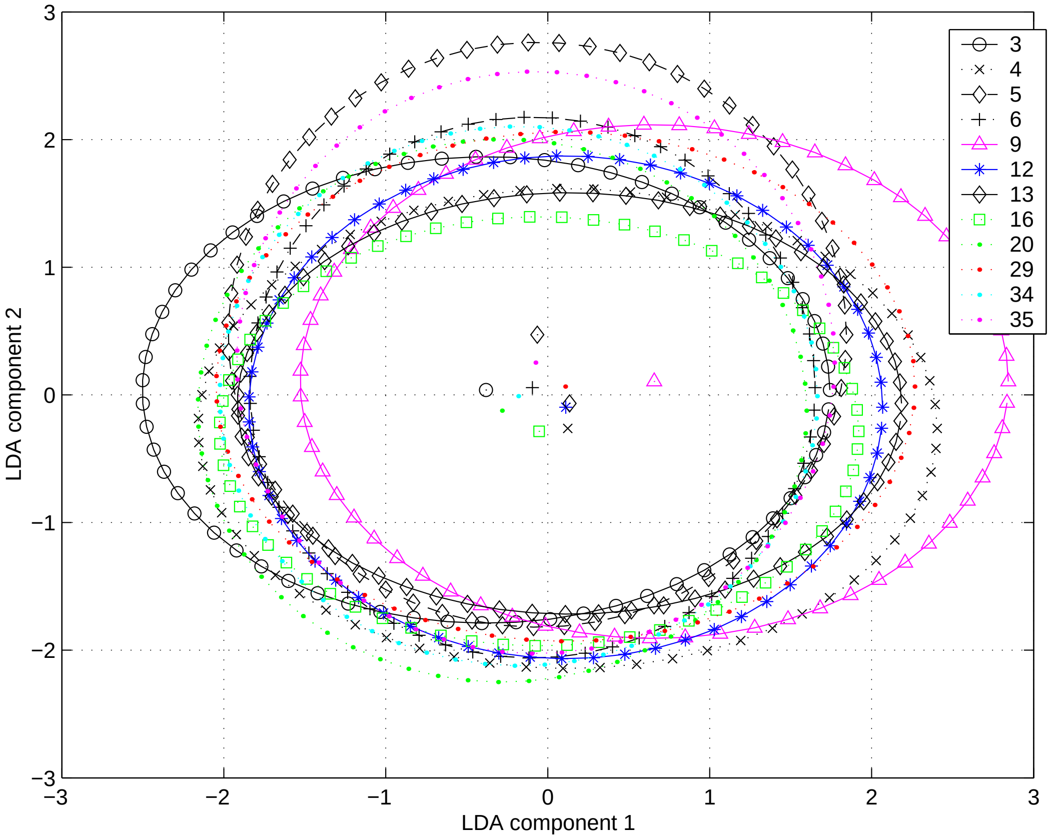

Figure 5: A plot of SSC with 16 centroids using the first two principal LDA components of 16 speakers from the XM2VTS database. Each number in the legend corresponds to an identity. The mean for each identity is represented as a dot while the first standard deviation along the two axes are shown as a circle centered on the mean.

the degree of separability of these two features. The first goal is achieved using Linear Discriminant Analysis (LDA) while the second goal is achieved using the F-Ratio test.

A Projection of Features onto the LDA Space

In this analysis using LDA, we extracted 16 centroids (for SSCs) and 12 MFCCs (using 24 filter-banks) and projected them onto the LDA space using all the eigen-vectors of $\Sigma_w^{-1}\Sigma_b$. The mean and standard deviation of the first two components of LDA are plotted in Figure 5 for SSCs and Figure 6 for MFCCs. An immediate observation about these figures is that the between-class variance is many more times smaller than the within-class variance. Another interesting aspect to observe is that speaker 3 and 6 have almost the same mean in the MFCC-induced LDA space (hence difficult to separate) but have different means in SSC-induced LDA space (hence separable). The same observation can be made for speaker 5 and 12 in the SSC-induced LDA space. Hence, speakers which are difficult to separate in one feature might stand a higher chance to be separable in another feature space. This complementary characteristic is what one seeks in order to improve the performance by some combination mechanisms.

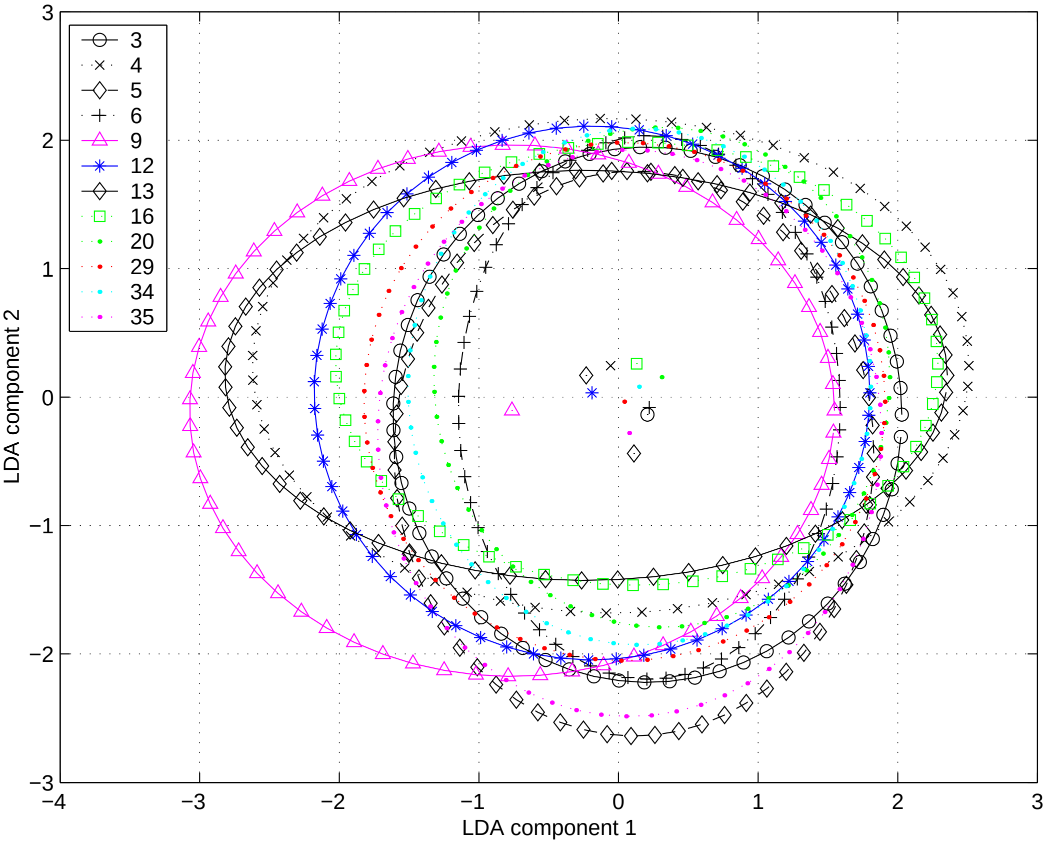

Figure 6: A plot of MFCC with 12 cepstrums using the first two principal LDA components of 16 speakers from the XM2VTS database

B Comparison of MFCCs and SSCs Using F-Ratio

We wish to compute the F-ratio statistics using $\frac{det|\Sigma_b|}{det|\Sigma_w|}$. The bigger this value is, the easier the features are to be classified. A preliminary study using this F-ratio on this dataset reveals that this value is indeed very small, when the covariance matrix dimension (due to the features) is large[4].

For practical reasons, we propose two solutions to compute the F-ratios: calculate the log of F-ratio, i.e., $\log\left(\frac{det|\Sigma_b|}{det|\Sigma_w|}\right)$; and use the matrix trace of the between- and within-class covariance matrix instead, i.e., $\frac{trace|\Sigma_b|}{trace|\Sigma_w|}$.

[4] The determinant of a matrix $\mathbf{x}$ is equivalent to $det|\mathbf{x}| = \prod_i \lambda_i$, where λ_i for $i = 1 \dots D$ are the eigen values of the matrix $\mathbf{x}$ of $D \times D$ dimensions. When $\lambda_i \to 0$, $\prod_i \lambda_i$ will approach zero while $\sum_i \lambda_i$ (which is the definition of the trace of $\mathbf{x}$, i.e, $trace(\mathbf{x})$) will not be influenced by small values of λ_i. This is especially true when the dimension D of the matrix $\mathbf{x}$ is very high.

Table 3: Different F-ratio criteria applied on the covariance matrices of MFCCs, PCA-transformed MFCCs, SSC and PCA-transformed SSCs.

Features	F-ratio criteria	
	$\log\left(\frac{det\|\Sigma_b\|}{det\|\Sigma_w\|}\right)$	$\frac{trace\|\Sigma_b\|}{trace\|\Sigma_w\|}$
MFCC	-101.961	0.0123
MFCC PCA (95%)	-57.366	0.0120
MFCC PCA (100%)	-101.714	0.0123
SSC	-260.109	0.0091
SSC PCA (95%)	-182.299	0.0091
SSC PCA (100%)	-107.241	0.0084

The calculating of $\log\left(\frac{det|\Sigma_b|}{det|\Sigma_w|}\right)$ can be performed as follows:

$$\begin{aligned}\log\left(\frac{det|\Sigma_b|}{det|\Sigma_w|}\right) &= \log(det|\Sigma_b|) - \log(det|\Sigma_w|) \\ &= \log\left(\prod_i \lambda_i^b\right) - \log\left(\prod_i \lambda_i^w\right) \\ &= \sum_i \log(\lambda_i^b) - \sum_i \log(\lambda_i^w)\end{aligned}$$

where λ_i^b and λ_i^w are the eigen values of the between- (Σ_b) and within-class (Σ_w) matrix, respectively.

In theory, the determinant of the *similarity transform* of a matrix is equal to the determinant of the original matrix [5, pp 311], i.e.,

$$\begin{aligned}det|\mathbf{BAB}^{-1}| &= det|\mathbf{B}|\ det|\mathbf{A}|\ det|\mathbf{B}^{-1}| \\ &= det|\mathbf{B}|\ det|\mathbf{A}|\ \frac{1}{det|\mathbf{B}|} = det|\mathbf{A}|\end{aligned}$$

The same property also applies to a matrix trace. Hence, we can literally take the features, transform them in another space (we use principal component analysis or PCA, in this case), calculate the covariance matrix and find its F-ratio. This F-ratio should be similar to the one calculated using the covariance matrix on the features directly. A simple experiment is performed on this dataset and the results are shown in Table 3. It can be seen that the F-ratios calculated using the determinant of the between- and within-class matrix are affected by small numbers while the F-ratios calculated using the matrix trace are not, i.e., the former F-ratios have greater differences than the ones using the latter, even though the log function is applied. In either case, the F-ratio of MFCCs is larger than that of SSCs. Hence, we expect that in terms of performance for the task of speaker authentication, MFCCs will be better than SSCs. Empirical results in Sections 3.5 and 4 confirm this observation.

References

[1] Y. Bengio. *Neural Networks for Speech and Sequence Recognition.* Thompson Computer Press, 1995.

[2] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1999.

[3] E. Chilton and H. Marvi. Two-Dimensional Root Cepstrum as Feature Extraction Method for Speech Recognition. *Electronics Letters*, 3(10):815–816, 2003.

[4] R. de Mori and M. Palakal. On the Use of a Taxonomy of Time-Frequency Morphologies for Automatic Speech Recognition. In *Int'l Joint Conf. Artificial Intelligence*, pages 877–879, 1985.

[5] G. H. Golub and C. F. Van Loan. *Matrix Computations.* 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[6] Astrid Hagen. *Robust Speech Recognition Based on Multi-Stream Processing.* PhD thesis, Ecole Polytechnique Federale de Lausanne, Switzerland, 2001.

[7] H. Hermansky, N. Morgan, Aruna Bayya, and Phil Kohn. Rasta-PLP speech analysis. In *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, volume 1, pages 121–124, San Francisco, 1992.

[8] S. S. Kajarekar and H. Hermansky. Analysis of Information in Speech and its Application in Speech Recognition. In *3rd Int'l Workshop Text, Speech and Dialogue (TSD'2000)*, pages 283–288, Brno, Czech Republic, September 2000.

[9] J. Lüttin. Evaluation Protocol for the XM2FDB Database (Lausanne Protocol). Communication 98-05, IDIAP, Martigny, Switzerland, 1998.

[10] I. Magrin-Chagnolleau, G. Gravier, M. Seck, O. Boeffard, R. Blouet, and F. Bimbot. A Further Investigation on Speech Features for Speaker Characterization. In *Proc. Int'l Conf. Spoken Language Processing*, volume 3, pages 1029–1032, Beijing, October 2000.

[11] A. Martin. NIST Year 2001 Speaker Recognition Evaluation Plan, 2001.

[12] K. K. Paliwal. Spectral Subband Centroids Features for Speech Recognition. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 617–620, Seattle, 1998.

[13] L. Rabiner and B-H Juang. *Fundamentals of Speech Recognition.* Oxford University Press, 1993.

[14] D. A. Reynolds. Experimental Evaluation of Features for Robust Speaker Identification. *IEEE Trans. Speech and Audio Processing*, 2(4):639–643, 1994.

[15] D. A. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1–3):19–41, 2000.

[16] M. K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling Dynamic Prosodic Variation for Speaker Verification. In *Proc. Int'l Conf. Spoken Language Processing*, volume 7, pages 3189–3192, Sydney, 1998.

[17] M. Kemal Sönmez, Larry Heck, Mitchel Weintraub, and Elizabeth Shriberg. A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. In *Proc. Eurospeech*, volume 3, pages 1291–1394, Rhodes, 1997. Greece.

[18] F. K. Soong and A.E. Rosenberg. On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, volume 36, pages 871–879, 1988.

[19] A. Varga and H. Steeneken. Assessment for Automatic Speech Recognition: NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. *Speech Communication*, 12(3):247–251, 1993.