Spectral Subband Centroids as Complementary Features for Speaker Authentication

Norman Poh Hoon Thian, Conrad Sanderson, and Samy Bengio

IDIAP, Rue du Simplon 4, CH-1920 Martigny, Switzerland
norman@idiap.ch, conradsand@ieee.org, bengio@idiap.ch

Abstract. Most conventional features used in speaker authentication are based on estimation of spectral envelopes in one way or another, 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 formant frequencies but are limited to a given subband. Empirical experiments carried out on the NIST2001 database using SSCs, MFCCs, LFCCs and their combinations by concatenation suggest that SSCs are somewhat more robust compared to conventional MFCC and LFCC features as well as being partially complementary.

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 is some literature on designing new and effective speech features for speaker authentication [8] (i.e., Line Spectrum Pairs, Time-Frequency Principal Component and Discriminant Components of the Spectrum), Mel-scale Frequency Cepstral Coefficients (MFCCs), which are commonly used in speech recognition, remain the state-of-the-art features, as far as speaker authentication is concerned. Empirical studies in [12] showed that Linear-scale Frequency Cepstral Coefficients (LFCCs) [11] achieve comparable performance to that of MFCCs [12, 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 (RASTA-PLP) [6] 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. In [2, Sec. 3.3], frequency and amplitude information are extracted from "spectral lines" [5]. 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 in 2D space. With quantisation on frequency and amplitude, this frequency/amplitude encoded data was classified using a feed-forward network and was shown to achieve a 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 models [16]. Follow-up work [15] used the (local variation of) pitch dynamics which contain speaker's intonation (speaking style). In both works, the resultant pitch system was combined with the cepstral feature-based system by summation of (log-)likelihood scores over the same utterance. They all showed improvement over the baseline system.

In the context of *speech recognition*, frequency information can be represented in the form of Spectral Subband Centroids (SSCs) [9], 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 [9] 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 [15, 16], 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 since they can be incorporated at the frame-level (and of course at the classifier-score level) while this is not possible in [15, 16]. 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, constituting an interesting research question.

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

2 Spectral Subband Centroids

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 [9], is defined as:

$$C_m = \frac{\int_{l_m}^{h_m} f w_m(f) P^\gamma(f) df}{\int_{l_m}^{h_m} w_m(f) P^\gamma(f) df} \qquad (1)$$

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. In the case of MFCCs, w_m is a triangular window. The same window is used here. The use of γ parameter in this function is a design parameter and is not motivated by any psychological aspect of hearing. The γ parameter has been used elsewhere in the literature [4] 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.

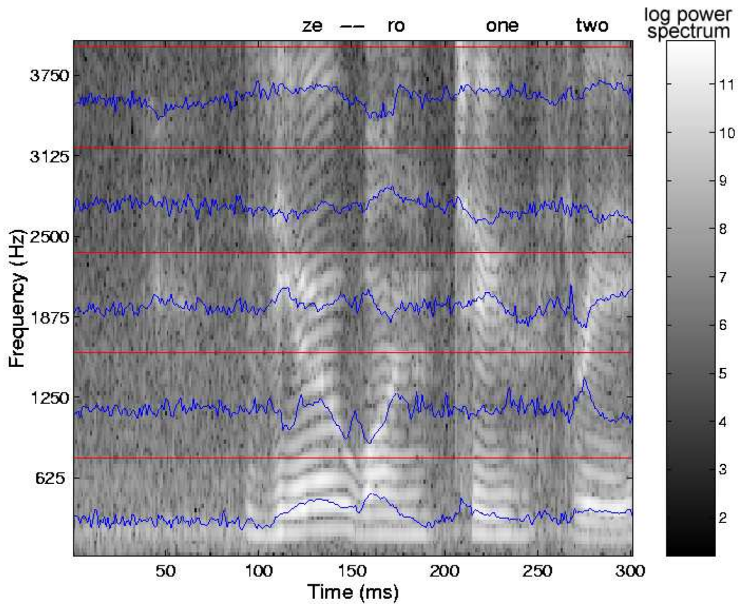

Fig. 1. SSC features across time

Figure 1 shows a conventional spectrogram overlaid with the SSC features with five equally-spaced bands, calculated using square windows. The 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 in that subband. This coincides with the idea of spectral lines [5] discussed earlier. However, in this context, the representation is limited to one value per subband. Secondly, if there is not enough centroids, then SSCs will not adequately represent a given speech signal.

Prior to testing SSCs using a real-life noisy database, we carried out several preliminary studies on SSCs using Linear Discriminant Analysis (LDA) under the Analysis of Variance (ANOVA) framework [10]. A subset of XM2VTS database and the female development set of NIST2001 (same as the one described in Section 3) were used for this test. The LDA analysis was used because it can separate useful sources of variance (e.g. physical articulatory features) from harmful sources of variance (e.g. handset differences, environmental and channel noise) [7]. We outline several conclusions of the preliminary studies reported in [10]:

– Based on LDA, we showed that about 12 to 16 centroids cover 99% of variance that is speaker discriminative. If less than 12 centroids are used, the speech utterance will be under-represented.
– Additional experiments based on LDA suggest that class labels (speaker’s identities) not separable in SSC feature space are separable in MFCC (Mel-scale Frequency Cepsstrum Coefficient) feature space. This suggests that SSCs are potentially complementary to MFCCs.
– The Fisher-ratio test showed 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.
– Preliminary empirical experiments on the female development subset of NIST2001 showed that about 16 to 18 centroids are optimal for speaker authentication.

– A theoretical study showed that mean-subtracted SSCs can somewhat reduce the effects of additive noise. The mean subtraction is done as follows:

$$C_m - E\{C_m\} \tag{2}$$

where $E\{C_m\}$ is the expectation of C_m over the whole utterance in a single access claim. The demonstration began with the assumption that a signal is composed of additive noise and the original clean signal. Deriving SSCs and mean-subtracted SSCs using this formulation, we showed that the additive component is partially cancelled during the mean subtraction. Empirical studies on NIST2001 also strongly supported this observation.
– Lastly, we showed empirically that first temporal derivates (deltas) of SSCs can also be used to further improve the performance.

The above studies were limited to studying the characteristics of SSCs compared to MFCCs under clean conditions. In this paper, the aspect of noise-robustness is evaluated.

3 Experiment Setup

In this study, a subset of NIST2001 was used to evaluate how well these features perform on telephone data with and without additive environmental noise, on speaker authentication tasks. It was obtained from the Switchboard-2 Phase 3 Corpus collected by the Linguistic Data Consortium. In this paper, only the female subset (which is known to be slightly more difficult than the male subset) was used for evaluation. In the original database, data for two different handsets are present (i.e., carbon and electret). However, only data from electret handsets were 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. There are 2694 accesses for the development population and 32029 accesses for the evaluation population, with a proportion of 10% of true claimant accesses. Four types of noise (`white`, `oproom` (for operational room), `factory` and `lynx`), taken from the NOISEX-92 database [17], were used to contaminate the NIST2001 dataset.

The classifier used in this paper is based on Gaussian Mixture Models (GMMs), similar to the one used in [13]. It models the statistical distribution of training feature vectors for each client. Briefly, a common impostor GMM model (also called a world model) is first obtained from the said 218 speakers using the Expectation-Maximization algorithm [3]. The world model is then adapted to each client's speech features using Maximum *a Posteriori* (MAP) estimation [13]. To make a decision, an average log-likelihood ratio between the client-adapted model and the world model (over all feature frames) is compared to a threshold chosen on development data.

The commonly used Half Total Error Rate (HTER) is used as evaluation criterion[1]. 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 $|\mathrm{FAR}(\theta) - \mathrm{FRR}(\theta)|$ is minimised with respect to θ. It is then used to make decisions on an evaluation set. Hence, the

[1] It should be noted the popular Equal Error Rate (EER) *was not used* here because this criterion does not reflect real applications where a threshold must be fixed in advance. Moreover, the use of DET or ROC curves to compare two systems has recently been shown to be erroneous and misleading [1], despite the fact that they are widely accepted in the literature.

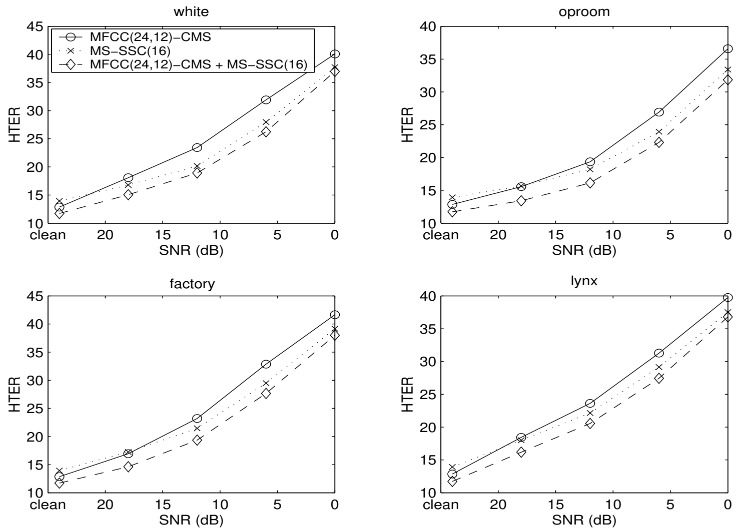

Fig. 2. *A priori* HTERs (in %) of SSCs, MFCCs and MFCC+SSC feature sets on the female evaluation subset of NIST2001 database, under mismatched conditions, using thresholds estimated on clean development data.

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 the HTER is, the better the performance.

4 Empirical Results in Mismatched Conditions

Preliminary studies in [10] showed that the following configuration of SSCs was optimal for the 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 [10]). Furthermore, only bands in the 300-3400 Hz frequency range are used. The log of delta energy is also used. To accomplish energy normalisation, the absolute log energy is not used.

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 improve performance. 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. 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 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 were used as classifiers for all experiments. The number of Gaussians was found by cross-validation based on the LFCCs features.

The results are shown in Figures 2 and 3 for these two sets of experiments.

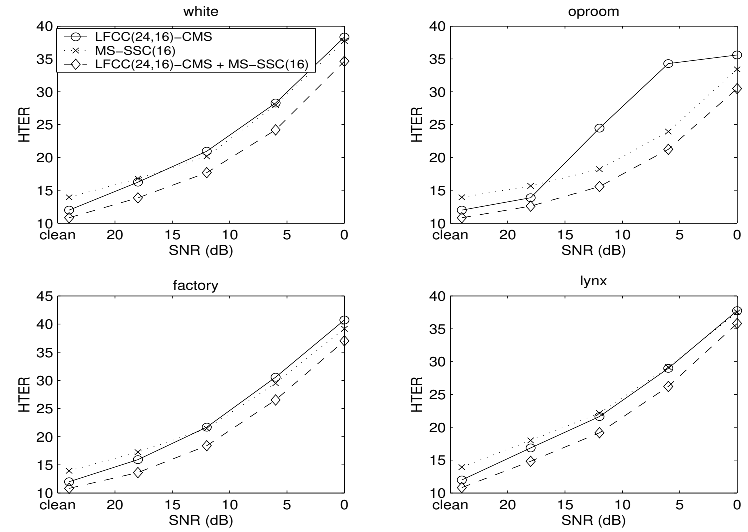

Fig. 3. *A priori* HTERs (in %) of SSCs, LFCCs and LFCC+SSC feature sets on the female evaluation subset of NIST2001 database, under mismatched conditions, using thresholds estimated on clean development data.

For both sets of experiments, it can be observed that MFCCs (respectively LFCCs) perform better than SSCs under clean conditions but are 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 relatively new 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. Preliminary findings in [10] based on ANOVA and LDA showed that SSCs are potential complementary features to conventional features such as MFCCs. In this paper, we valided these findings using the female development subset of the NIST2001 SwitchBoard database. Based on the results, it is concluded that 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. Hence, dominant frequencies represented by SSCs contain speaker discriminative information, somewhat different from what MFCCs (respectively LFCCs) provide. One potential future direction to study the usefulness of the medium to long-term time-trajectory of SSCs. This is motivated by [15], where it is shown that speaker's pitch dynamics (speaker's intonation) are useful for speaker authentication. The advantage of using the time-trajectory of SSCs as compared to pitch dynamics is that not only that the (low frequency) pitch is included, the whole frequency band is actually taken into account.

6 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.

References

1. S. Bengio, M. Keller, and J. Mariéthoz. The Expected Performance Curve. IDIAP Research Report 03-85, Martigny, Switzerland, 2003.
2. Y. Bengio. *Neural Networks for Speech and Sequence Recognition*. Thompson Computer Press, 1995.
3. C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.
4. E. Chilton and H. Marvi. Two-Dimensional Root Cepstrum as Feature Extraction Method for Speech Recognition. *Electronics Letters*, 3(10):815–816, 2003.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. N. Poh, C. Sanderson, and S. Bengio. An Investigation of Spectral Subband Centroids For Speaker Authentication. IDIAP Research Report 03-62, Martigny, Switzerland, 2003. to appear in Int'l Conf. on Biometric Authentication, Hong Kong, 2004.
11. L. Rabiner and B-H Juang. *Fundamentals of Speech Recognition*. Oxford University Press, 1993.
12. D. A. Reynolds. Experimental Evaluation of Features for Robust Speaker Identification. *IEEE Trans. Speech and Audio Processing*, 2(4):639–643, 1994.
13. D. A. Reynolds, T. Quatieri, and R. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. 10(1–3):19–41, 2000.
14. C. Sanderson. Speech Processing & Text-Independent Automatic Person Verification. IDIAP Communication 02-08, Martigny, Switzerland, 2002.
15. 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.
16. 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.
17. 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.