Speaker Verification with Adaptive Spectral Subband Centroids

Tomi Kinnunen[1], Bingjun Zhang[2], Jia Zhu[2], and Ye Wang[2]

[1] Speech and Dialogue Processing Lab
Institution for Infocomm Research (I^2R)
21 Heng Mui Keng Terrace, Singapore 119613
ktomi@i2r.a-star.edu.sg
[2] Department of Computer Science
School of Computing, National University of Singapore (NUS)
3 Science Drive 2, Singapore 117543
{bingjun,zhujia,wangye}@comp.nus.edu.sg

Abstract. Spectral subband centroids (SSC) have been used as an additional feature to cepstral coefficients in speech and speaker recognition. SSCs are computed as the centroid frequencies of subbands and they capture the dominant frequencies of the short-term spectrum. In the baseline SSC method, the subband filters are pre-specified. To allow better adaptation to formant movements and other dynamic phenomena, we propose to adapt the subband filter boundaries on a frame-by-frame basis using a globally optimal scalar quantization scheme. The method has only one control parameter, the number of subbands. Speaker verification results on the NIST 2001 task indicate that the selection of the parameter is not critical and that the method does not require additional feature normalization.

1 Introduction

The so-called *mel-frequency cepstral coefficients* [1] (MFCC) have proven to be efficient feature set for speaker recognition. A known problem of cepstral features, however, is noise sensitivity. For instance, convolutive noise shifts the mean value of the cepstral distribution whereas additive noise tends to modify the variances [2]. To compensate for the feature mismatch between training and verification utterances, normalizations in feature, model and score domains are commonly used [3].

Spectral subband centroids [4; 5; 6; 7] (SSC) are an alternative to cepstral coefficients. SSCs are computed as the centroid frequencies of subband spectra and they give the locations of the local maxima of the power spectrum. SSCs have been used for speech recognition [4; 5], speaker recognition [7] and audio fingerprinting [6]. Recognition accuracy of SSCs is lower in noise-free conditions compared with MFCCs. However, SSCs can outperform MFCCs in noisy conditions and they can be combined with MFCCs to provide complementary information.

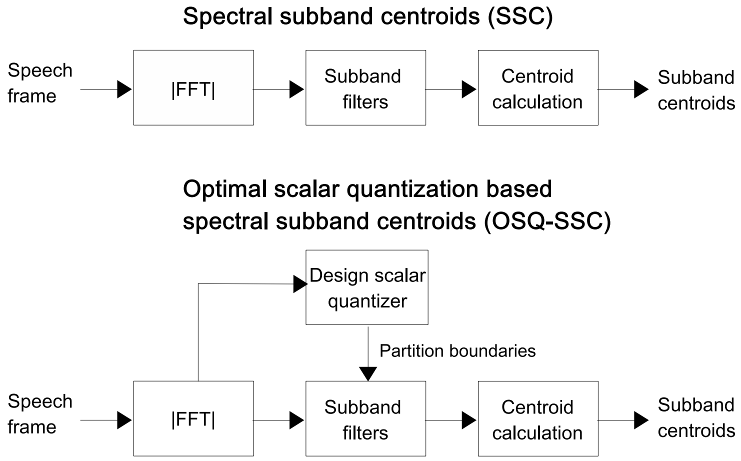

Fig. 1. Computation of the SSC and the proposed OSQ-SSC features. In SSC, the subband boundaries are fixed and in OSQ-SSC, the boundaries are re-calculated for every frame by partitioning the spectrum with optimal scalar quantization.

The key component of the SSC method is the filterbank. Design issues include the number of subbands, the cutoff frequencies of the subband filters, the shape of the subband filters, overlapping of the subband filters, compression of the spectral dynamic range and so on [5]. The parameters of the filterbank can be optimized experimentally for a given task and operating conditions.

In this study, our aim is to simplify the parameter setting of the SSC method by adding some self-adaptivity to the filterbank. In particular, we optimize the subband filter cutoff frequencies on a frame-by-frame basis to allow better adaptation to formant movements and other dynamic phenomena. We consider the subbands as *partitions* or *quantization cells* of a *scalar quantizer*. Each subband centroid is viewed as the representative value of that cell and the problem can be defined as joint optimization of the partitions and the centroids. The difference between the conventional SSC method and the proposed method is illustrated in Fig. 1.

2 Spectral Subband Centroids

In the following, we denote the FFT magnitude spectrum of a frame by $S[k]$, where $k = 1, \ldots, N$ denotes the discrete frequency index. The index $k = N$ corresponds to the half sample rate $f_s/2$. The m^{th} subband centroid is computed as follows [5]:

$$c_m = \frac{\sum_{k=q_l(m)}^{q_h(m)} k W_m[k] S^\gamma[k]}{\sum_{k=q_l(m)}^{q_h(m)} W_m[k] S^\gamma[k]}, \tag{1}$$

where $W_m[k]$ are the m^{th} bandpass filter coefficients, $q_l(m), q_h(m) \in [1, N]$ are its lower and higher cutoff frequencies and γ is a dynamic range parameter.

The shape of the subband filter introduces bias to the centroids. For instance, the triangular shaped filters used in MFCC computation [1] shift the centroid towards the mid part of the subband. To avoid such bias, we use a uniform filter in (1): $W_m[k] = 1$ for $q_l(m) \leq k \leq q_h(m)$. Furthermore, we set $\gamma = 1$ in this study. With these modifications, (1) simplifies to

$$c_m = \frac{\sum_{k=q_l(m)}^{q_h(m)} kS[k]}{\sum_{k=q_l(m)}^{q_h(m)} S[k]}. \tag{2}$$

3 Adapting the Subband Boundaries

To allow better adaptation of the subband centroids to formant movements and other dynamic phenomena, we optimize the filter cutoff frequencies on a frame-by-frame basis. We use scalar quantization as a tool to partition the magnitude spectrum into K non-overlapping quantization cells. The subband cutoff frequencies, therefore, are given by the partition boundaries of the scalar quantizer.

The expected value of the squared distortion for the m^{th} cell is defined as

$$e_m^2 = \sum_{q(m-1)<k\leq q(m)} p_k (k - c_m)^2, \tag{3}$$

where $p_k = S[k] / \sum_{n=1}^{N} S[n]$ is the normalized FFT magnitude, c_m is the subband centroid as defined in (2) and $q(m-1), q(m)$ are the subband boundaries: $0 = q(0) < q(1) < q(2) < \cdots < q(K) = N$. The scalar quantizer design can then be defined as the minimization of the total error:

$$\min_{(q(0),q(1),\ldots,q(K))} \sum_{m=1}^{K} e_m^2. \tag{4}$$

The number of subbands (K) is considered as a control parameter that needs to be optimized experimentally for a given application.

We have implemented a globally optimal scalar quantizer which uses matrix searching technique to solve (4) [8]. The time complexity of the method is $O(KN)$ and our implementation runs 18 times faster than realtime on a 3 GHz Pentium processor for $(K, N) = (8, 128)$. It is interesting to note that optimal algorithms for vector quantization [9] require exponential time but globally optimal scalar quantizer can designed in polynomial time. This theoretically interesting property, in fact, was one of our initial motivations to apply the method to feature extraction. We term the proposed method as *optimal scalar quantization based spectral subband centroids* (OSQ-SSC).

Figure 2 shows the centroids from both the SSC with mel filterbank and the OSQ-SSC method. The spectrogram is also shown as a reference. It can be seen that the OSQ-SSC features are better adapted to local dynamic changes of the spectrum compared with SSC. In particular, the centroids from OSQ-SSC tend to follow the F0 harmonics and the formant frequencies during voiced regions.

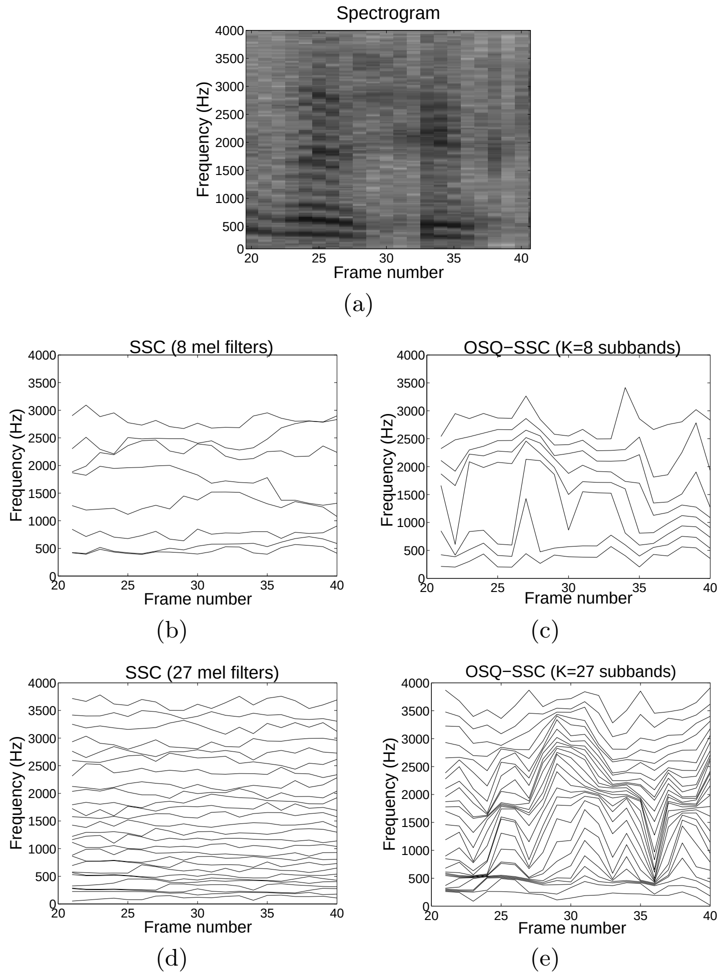

Fig. 2. Illustrations of conventional spectral subband centroids with fixed filterbank (SSC) and the proposed method with adapted subband boundaries (OSQ-SSC).

4 Speaker Verification Setup

We use the NIST 2001 speaker recognition evaluation corpus, a conversational cellular phone corpus, in our experiments[3]. The 1-speaker detection task as defined by NIST consists of 174 target speakers and 22418 verification trials with an genuine/impostor ratio of 1:10. The amount of training data per speaker is about 2 minutes and the duration of the test segments varies from a few seconds up to one minute.

[3] `http://www.nist.gov/speech/tests/spk/`

We use the Gaussian mixture model - universal background model (GMM-UBM) with diagonal covariance matrices as the recognizer [10]. The background model is trained from the development set of the NIST 2001 corpus using the expectation-maximization (EM) algorithm. Target speaker models are derived with maximum a posteriori (MAP) adaptation of the mean vectors and the verification score is computed as the average log-likelihood ratio. The GMMs have 256 Gaussian components for all the features and parameters tested.

We include the standard MFCC front-end as a reference, including 12 MFCC with their delta and double-delta coefficients. RASTA filtering, energy-based voice activity detection (VAD) and mean/variance normalization are applied to enhance robustness. The same VAD is used with SSC and OSQ-SSC features.

5 Speaker Verification Results

We first study the number of subbands and feature normalization. Feature normalization is performed after voice activity detection to give zero mean and unit variance features. The equal error rates (EER) for the NIST 2001 evaluation set are shown in Table 1. The best result is obtained without any normalization, indicating that OSQ-SSC is a robust feature in itself (In contrast, both the mean and variance normalization were helpful for MFCC). Optimal number of subbands is $K = 8$. For too few subbands, speaker discrimination is expected to be poor. On the other hand, for too many subbands, the variance of the centroids becomes less, removing useful information as well.

Table 1. Effects of #subbands and normalization to OSQ-SSC accuracy (EER %)

Subbands	Feature normalization			
	None	Mean	Var	Mean+Var
4	19.4	27.5	31.7	27.5
8	18.0	24.8	29.8	24.5
16	19.1	23.1	27.5	23.3
32	19.8	24.6	28.4	24.2

Table 2. Comparison of static and delta features of OSQ-SSC (EER %).

Subbands	Feature set		
	OSQ-SSC	ΔOSQ-SSC	OSQ-SSC + Δ
4	19.4	26.9	20.3
8	18.0	22.2	19.7
16	19.1	23.4	20.3
32	19.8	24.0	21.3

We next compare OSQ-SSC, their delta coefficients and the concatenation of the static and delta coefficients at the frame level. Based on Table 1, we turn off the normalizations. The results are given in Table 2. The delta coefficients yield higher error rates compared with the static coefficients which is expected. We did, however, expect some improvement when combining static and delta features which is not the case. The error rates of the delta coefficient are relatively high compared with the static coefficient which partly explains why the fusion is not successful. The simple differentiator method for computing deltas may not be robust enough and other methods like linear regression should be considered.

Table 3. Comparison of MFCC, OSQ-SSC and ΔOSQ-SSC under additive white noise condition (EER %).

Noise weight (α)	Feature set MFCC	OSQ-SSC	ΔOSQ-SSC
0	8.3	18.0	22.2
0.3	15.7	19.9	26.6
0.6	18.4	22.6	29.2
0.9	25.6	28.7	47.9

We next study noise robustness of the OSQ-SSC feature. We contaminated all the training and testing files with additive white noise with three different noise levels. The noise was added with linear combination of the speech and noise as $x_{\text{noisy}}[n] = \alpha z[n] + (1-\alpha)x_{\text{orig}}[n]$, where $x_{\text{noisy}}[n]$, $z[n]$ and $x_{\text{orig}}[n]$ denote the noisy speech, noise and the original speech, respectively. The results for $K = 8$ subbands and their delta coefficients are given in Table 3. The MFCC result is shown as a reference.

All the three features degrade when noise level is increased as expected. The MFCC feature gives the best result in all cases and ΔOSQ-SSC gives the worst result in all conditions. However, relative degradation of OSQ-SSC is much smaller compared with MFCC. For instance, the relative increase in EER from $\alpha = 0$ to $\alpha = 0.3$ is 89 % for MFCC, whereas for OSQ-SSC it is only 11 %. This is interesting since the MFCC features have 36 dimensions, including deltas and double deltas, mean and variance normalization and RASTA filtering. In turn, OSQ-SSC has only 8 dimensions and is without any normalizations. We interpret the result so that the intrinsic resistance to additive noise of OSQ-SSC is better than that of MFCC. On the other hand, speaker discrimination of MFCC is clearly higher.

Finally, we compare OSQ-SSC with SSC. We consider the following three filterbank configurations:

- SSC(1) : linear frequency scale, non-overlapping rectangular filters
- SSC(2) : mel frequency scale, non-overlapping rectangular filters
- SSC(3) : mel frequency scale, overlapping triangular filters

According to [7], mean subtraction helps SSC. We confirmed this experimentally and we apply it in all the three cases. The results are shown in Table 4.

Table 4. Comparison SSC and OSQ-SSC (EER %).

Subbands	Feature set			
	OSQ-SSC	SSC(1)	SSC(2)	SSC(3)
4	19.4	24.8	23.9	24.8
8	18.0	19.7	21.9	15.4
16	19.1	21.0	25.3	17.5
32	19.8	24.2	26.3	22.5

The performance of the SSC method strongly depends on the parameter setting. The best SSC result (EER=15.4 %) is obtained by using eight overlapping filters on the mel-scale. Overall, SSC(3) gives the best result among the three filterbank configurations, followed by SSC(1) and SSC(2), respectively. Overlapping filters are useful for SSC.

Comparing OSQ-SSC with SSC, OSQ-SSC is less sensitive to parameter setup. The method has only one control parameter and the results incidate that the method is not sensitive to it. For SSC(3), the error rate varies between 15.4% - 24.8 % whereas for OSQ-SSC, the range is 18.0 % - 19.8 %. The OSQ-SSC method has a built-in “self-optimizing” property of the filterbank. The success of the SSC method, on the other hand, is expected to depend much on the correct setting of the filterbank parameters.

6 Conclusions

We have simplified the spectral subband centroid (SSC) method by adding self-adaptivity to the filterbank. The proposed method (OSQ-SSC) has only one control parameter and the experiments indicated that the method is not very sensitive to it. It was also found that the proposed method does not require normalization which help both the MFCC and the baseline SSC features. This is beneficial for real-time applications.

The advantage of OSQ-SSC is that the subband boundaries are adapted for each frame. The partitions of the scalar quantizer, however, are non-overlapping which limits the centroid movements. According to our experiments, overlapping filters would be beneficial to the SSC method. An interesting future direction, therefore, would be studying optimization of the filterbank boundaries with overlapping filters.

Overall, accuracy of the MFCC features is better compared with the SSC-related features. It would be interesting to study the usefulness of the OSQ-SSC features for other applications like speech recognition. It is fairly possible that the centroids from OSQ-SSC capture formant-like features which are related to speech content rather than the speaker.

Bibliography

[1] Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech, and Signal Processing **28**(4) (1980) 357–366

[2] Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001), Crete, Greece (2001) 213–218

[3] Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.: A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing **2004**(4) (2004) 430–451

[4] Gajić, B., Paliwal, K.: Robust speech recognition in noisy environments based on subband spectral centroid histograms. IEEE Trans. Audio, Speech and Language Processing **14**(2) (2006) 600–608

[5] Paliwal, K.: Spectral subband centroid features for speech recognition. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1998). Volume 2., Seattle, USA (1998) 617–620

[6] Seo, J., Jin, M., Lee, S., Jang, D., Lee, S., Yoo, C.: Audio fingerprinting based on normalized spectral subband centroids. In: Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005). Volume 3. (2005) 213–216

[7] Thian, N., Sanderson, C., Bengio, S.: Spectral subband centroids as complementary features for speaker authentication. In: Proc. First Int. Conf. Biometric Authentication (ICBA 2004), Hong Kong, China (2004) 631–639

[8] Wu, X.: Optimal quantization by matrix searching. Journal of Algorithms **12**(4) (1991) 663–673

[9] Gersho, A., Gray, R.: Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston (1991)

[10] Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1) (2000) 19–41