Dimension Reduction of the Modulation Spectrogram for Speaker Verification

Tomi Kinnunen

Speech and Image Processing Unit
Department of Computer Science
University of Joensuu, Finland
tkinnu@cs.joensuu.fi

Kong-Aik Lee and Haizhou Li

Speech and Dialogue Processing Lab
Human Language Technology Department
Institute for Infocomm Research (I^{2}R), Singapore
{kalee,hli}@i2r.a-star.edu.sg

Abstract

A so-called modulation spectrogram is obtained from the conventional speech spectrogram by short-term spectral analysis along the temporal trajectories of the frequency bins. In its original definition, the modulation spectrogram is a high-dimensional representation and it is not clear how to extract features from it. In this paper, we define a low-dimensional feature which captures the shape of the modulation spectra. The recognition accuracy of the modulation spectrogram based classifier is improved from our previous result of EER=25.1% to EER=17.4% on the NIST 2001 speaker recognition task.

Index Terms: modulation spectrum, spectro-temporal features, speaker recognition

1. Introduction

The human auditory system integrates information over an interval of several hundreds of milliseconds [1]. In speech processing, relevance of longer time context for phonetic classification has been supported by both information-theoretic analysis [2] as well as improvements in speech recognition through spectro-temporal features; for an overview, refer to [3].

In modern speaker recognition systems, on the other hand, contextual and long-term information are extracted in a rather different way. First, the input utterance is converted into a sequence of tokens such as phone labels [4] or Gaussian mixture model (GMM) tokens [5]. This is followed by modeling of the token sequences using N-grams and support vector machines [4]. While these approaches have shown promising results especially when combined with traditional spectral features, their implementation is complex, and computational complexities high relative to the benefit obtained in the final recognition system. It is likely that the tokenizers also quantize the signal too much by losing some useful spectro-temporal details that could be useful for speaker recognition. These reasons have motivated us to study low-complexity contextual acoustic features, similar to those used in speech recognition, that incorporate contextual information directly to the feature coefficients [6, 7].

Our contextual features are based on the concept of the so-called *modulation spectrum* [1, 8]. Modulation spectrum is defined as the spectral representation of a temporal trajectory of a feature and it provides information of the dynamic characteristics of the signal. The modulation spectrum of a typical speech signal has a steep low-pass shape with most of the energy concentrated on modulation frequencies less than 20 Hz. Low-frequency modulations of the signal energy are related to speech rhythm which we hope to capture with the modulation spectrum-based features. As an example, it has been reported that conversational speech has a dominant modulation frequency component around 4 Hz which is roughly the same as the average syllable rate [1]. The dominant peak of the modulation spectrum, therefore, may be an acoustic correlate of speaking rate. Furthermore, some "high-level" speech phenomena can be seen as acoustic events characterized by low modulation frequencies. For instance, *laughter* consists of a few successive vowel-like bursts having similar spectral structure between the bursts and spaced equidistant in time. Thus, laughter is characterized by its "fundamental frequency" which may be a speaker-specific feature. These ideas motivate us to study the usefulness of the modulation spectrum for capturing some articulatory and stylistic features to be used in speaker recognition.

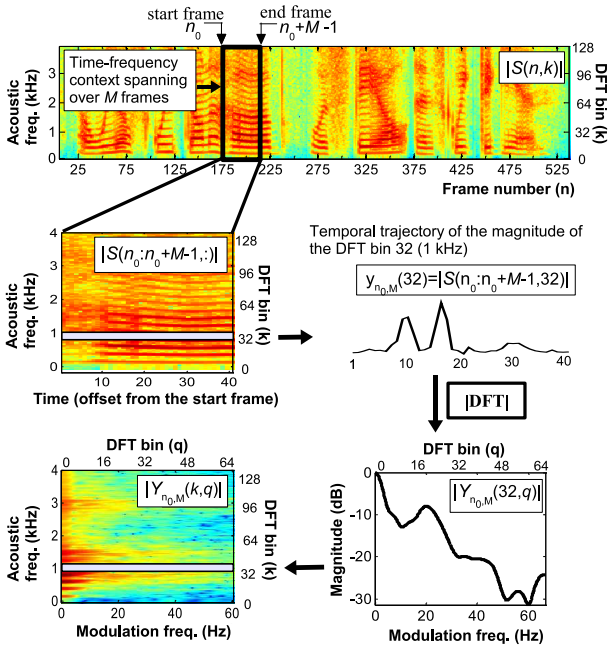

Figure 1: Computation of the modulation spectrogram from a spectrogram. A time-frequency context is extracted, from which the DFT magnitude spectra of all frequency bands are computed. We have used log magnitude values to improve visual appearance; however, all the computations use linear magnitude values.

A *joint acoustic and modulation frequency* representation [8] is obtained by simultaneous spectral analysis of all the frequency bins as illustrated in Fig. 1. This representation is

also known as the *modulation spectrogram* [9] and we will use this terminology for brevity. In [6], we presented preliminary speaker verification results by using the modulation spectrogram with a long-term averaging classifier equipped with divergence-based distance measure. Recently, modulation spectrogram has also been applied successfully to speaker separation in a single-channel audio by filtering in this domain [9].

In this paper, our primary goal is to explore the relative importances of the acoustic and modulation frequency resolutions and the effect of the time-frequency context length to speaker verification accuracy. In this way, we aim at establishing a reasonable baseline system for the modulation spectrogram based speaker verification. Short-term features like MFCCs have been largely studied whereas literature on long-term features is limited. Another motivation comes from the observation that modulation spectrum *filtering* has already been applied in the conventional speaker recognition systems via RASTA processing and computation of the delta coefficients of cepstral features [10, 11]. By studying the modulation spectrum as a *feature* in speaker verification, we aim at gaining more insight about the significance of the modulation spectrum *per se* for speaker verification. As a secondary goal, we wish to explore how the modulation spectrum based feature set compares with the standard MFCCs, and whether these two feature sets have fusion potential.

In our preliminary proposal [6], we restricted our recognition experiments to a simple long-term averaging classifier, followed by score normalization. The reason was that, in its original definition, the modulation spectrogram is a high-dimensional representation for which statistical models like GMM cannot be trained due to numerical problems (ill-defined covariance matrices) that are a result of the high dimensionality. In this study, we therefore define a lower-dimensional feature which represents the shape of the joint frequency representation. This lower-order approximation is achieved by using mel-frequency filtering on the acoustic spectrum and discrete cosine transform on the modulation spectrum. In this way, we are able to replace the averaging classifier with a standard Gaussian mixture model [12] recognizer and report updated recognition results.

2. The Modulation Spectrogram

2.1. Computing the Modulation Spectrogram

The modulation spectrogram is derived from the conventional spectrogram shown in the top panel of Fig. 1. To compute the spectrogram [13], the signal $s(n)$ is first divided into frames of length L samples with some overlapping between the successive frames. Each frame is pre-emphasized and multiplied by a Hamming window, followed by K-point DFT computation. The magnitude are retained which yields the magnitude spectrogram $|S(n,k)|$, where n denotes the frame index and k denotes the DFT bin ($0 \leq k \leq K/2$).

To derive the modulation spectrogram, the magnitude spectrogram is analyzed in short-term frames with some overlap, similar to the first transformation. Now the "frames", in fact, correspond to two-dimensional time-frequency contexts shown in the central left panel of Fig. 1. A time-frequency context, starting from frame n_0 and having length of M frames, consists of all the frequency bands within the time interval $[n_0, n_0+M-1]$. The temporal trajectory of the kth frequency band within the time-frequency context, denoted by $y_{n_0,M}(k)$, is therefore $y_{n_0,M}(k) = (|S(n_0,k)|, |S(n_0+1,k)|, \ldots, |S(n_0+M-1,k)|)$.

The modulation spectrum of the kth frequency bin is computed by multiplying $y_{n_0,M}(k)$ with a Hamming window and computing Q-point DFT. The magnitude of the DFT is retained, resulting in the modulation spectrum $|Y_{n_0,M}(k,q)|$. In summary, here k and q are the "acoustic" and "modulation" frequency indices, respectively, where $0 \leq k \leq K/2$ and $0 \leq q \leq Q/2$.

It should be noted that modulation spectrum can be computed also by convolving the original signal with a set of bandpass filter kernels, followed by some form of envelope detection. We have chosen the FFT-based method because it is straightforward to implement and computationally efficient.

2.2. Setting the Parameters

The most crucial parameters for the modulation spectrogram are the frame shift and the time-frequency context length. The frame shift determines the sampling rate of the temporal trajectories and hence sets the upper limit for the modulation frequencies. For instance, a typical frame shift of 20 milliseconds implies a modulation spectrum sampled at $1000/20 = 50$ Hz, and therefore, the highest modulation frequency is 25 Hz. For more details on the sampling considerations, refer to [14].

The time-frequency context length (M), on the other hand, is responsible for controlling the frequency resolution of the modulation spectrum. For a large M, the frequency resolution can be increased. However, for accurate spectrum estimation, M should be short enough so that the temporal trajectories remain stationary within the context. In our previous studies with temporal features, best verification results on the NIST corpora were obtained by using a time-frequency context of 200 to 300 milliseconds in length [6, 7]. Similar time-frequency contexts have been used in speech recognition [1, 3].

3. Reducing the Dimensionality

When used as a feature for speaker recognition, we rearrange the two-dimensional matrix $|Y_{n_0,M}(k,q)|$, where $0 \leq k \leq K/2$ and $0 \leq q \leq Q/2$, into a single vector of dimensionality $(K/2+1)(Q/2+1)$. For instance, for the typical values $K = 256$ and $Q = 128$, dimensionality is 8385. This is about two orders of magnitude too high to be used with statistical classifiers on typical speech training sample sizes. In principle, we can reduce K and Q by using a shorter frame and shorter context, respectively. This, however, leads to significant reductions in the respective frequency resolutions and also violates the idea of the contextual features expanding over a long time window. We prefer to keep the context size up to several hundreds of milliseconds and reduce the dimensionality of these features.

3.1. Reducing the Acoustic Frequency Dimension

We reduce the dimensionality of the acoustic frequency variable using a standard mel-frequency filterbank [13] which effectively reduces correlations between the frequency subbands. The standard triangular bandpass filters are applied on the short-term spectra and the temporal trajectories of the filter outputs are then subjected to modulation frequency analysis as described in the previous section.

We compared linear-frequency and mel-frequency filterbanks in preliminary experiments. It was found out that the mel-frequency filterbank outperforms the linear-frequency filterbank systematically, except for a small number of filters (5-10) for which the linear-frequency filterbank was slightly bet-

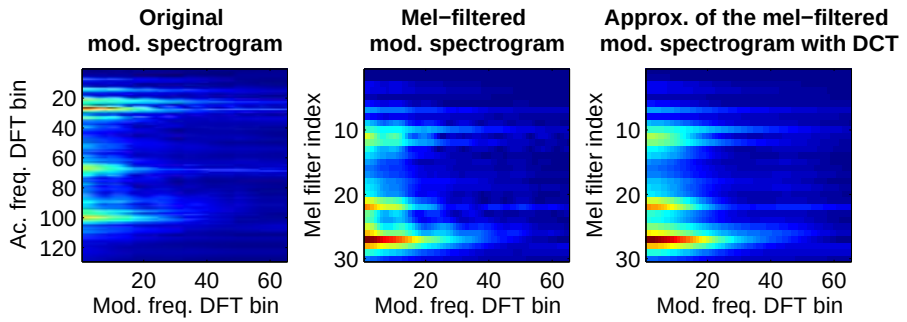

Figure 2: Dimension reduction of a single modulation spectrogram frame. Dimension reduction is achieved by mel-frequency filtering along the acoustic frequency axis (number of filters $C = 30$) and DCT compression along the modulation frequency axis (number of DCT coefficients $D = 4$). The dimensionalities of the features corresponding to the three panels are $129 \times 65 = 8385$, $30 \times 65 = 1950$ and $30 \times 4 = 120$, respectively.

ter. However, the performance increases when using more filters and therefore the mel-scale seems a better choice.

3.2. Reducing the Modulation Frequency Dimension

The modulation spectrum of an arbitrary frequency band contains redundant information as well. In particular, the modulation spectrum has a lowpass shape with a heavy damping of the frequencies above 20 Hz or so, and the spectrum is relatively smooth in shape. This suggests capturing the *envelope* of the modulation spectrum by using the *discrete cosine transform* (DCT), similar to cepstrum computation. We apply the DCT to each modulation spectrum, yielding a Q-dimensional vector of the DCT coefficients. We retain the lowest D coefficients, including the DC coefficient, so as to preserve most of the signal energy. To this end, by using C mel-frequency filters and retaining the lowest D cosine transformation coefficients, the feature vectors have dimensionality $C \times D$. Typical values are $C = 20$ and $D = 3$, implying vectors of dimensionality 60.

Figure 2 illustrates dimension reduction for a single matrix $|Y_{n_0,M}(k, q)|$. The panel on the left shows the original modulation spectra (linear frequency scale). The middle panel shows the mel-frequency modulation spectrum obtained using $C = 30$ filters and the panel on the right shows its approximation using $D = 4$ cosine transform coefficients. The approximation was produced by retaining the lowest 4 coefficients, followed by inverse DCT. It can be seen that the overall shape of the mel-frequency modulation spectrum is well retained. The dimensionalities of the features corresponding to the three panels are $129 \times 65 = 8385$, $30 \times 65 = 1950$ and $30 \times 4 = 120$, respectively.

3.3. Further Considerations

Nonlinear operators are commonly used in speech front-ends and we studied some of them in preliminary experiments as well. In particular, we experimented with (1) squaring of the FFT spectrum magnitude prior to mel-filtering, (2) log-compression of the mel-filter outputs prior to modulation frequency analysis and (3) log-compression of the modulation spectra prior to the final DCT. The first two nonlinearities yielded systematically higher error rates whereas the third one did not make significant change. While we do not have theoretical justifications for these results, based on our experiments, we recommend to use the simple magnitude operators without squaring or log-compression.

It is worth noting that the proposed feature includes similar operations to MFCC computation, but it does *not* reduce to MFCC vector when the context length is one frame ($M = 1$). In MFCC computation, the DCT is applied on the acoustic magnitude spectrum whereas we apply it to the modulation magnitude spectrum. It is easy to show that for $M = 1$, the proposed feature equals mel-filtered magnitude spectrum, but without log-compression and DCT as in MFCC.

4. Experimental Setup

We use the NIST 2001 speaker recognition evaluation corpus for our experiments. The NIST corpus consists of conversational telephony speech in English. The speech is recorded over the cellular telephone network with a sampling frequency of 8 kHz. We study the performance on the 1-speaker detection task which consists of 174 target speakers and a total number of 22,418 verification trials of which 90 % are impostor trials and 10 % are genuine speaker trials. The amount of training data is two minutes per speaker and the length of the test segment varies from a few seconds up to one minute.

The feature extraction parameters for the spectrogram were set as shown in Table 1, and these were kept fixed throughout the experiments while varying the modulation spectrogram parameters. We use the *Gaussian mixture model-universal background model* (GMM-UBM) with diagonal covariance matrices as the recognizer [12]. The UBM is trained using the development set of the NIST 2001 with the expectation-maximization (EM) algorithm. Target speaker models are derived using *maximum a posteriori* (MAP) adaptation of the mean vectors, and the verification score is computed as the average log-likelihood ratio. Speaker verification accuracy is measured in equal error rate (EER), which corresponds to the verification threshold at which the probabilities of false acceptance and false rejection are equal.

Table 1: Parameter setup of the spectrogram.

Spectrogram parameters	
Frame length	$L = 240$ samples (30 ms)
Frame shift	$(1/4)L = 60$ samples (7.5 ms)
Window function	Hamming
Pre-emphasis filter	$H(z) = 1 - 0.97z^{-1}$
FFT order	$K = 256$

Table 2: Effects of mel filtering and DCT to recognition accuracy (EER %).

Mel filters (C)	DCT coeffs. (D) 1	2	3	4
5	26.7	26.4	25.9	26.2
10	22.5	22.7	22.3	22.5
15	21.1	21.0	20.5	20.3
20	20.5	20.1	20.1	20.3
25	20.7	20.3	20.2	20.4
30	20.1	20.1	**19.9**	28.7
35	21.0	20.4	21.6	41.0
40	21.1	21.3	27.5	47.5

Table 3: Effects of the mel filtering and DCT compression to recognition accuracy (EER %) when keeping the dimensionality ($C \times D$) fixed to 60.

C	D	EER	C	D	EER
1	60	36.9	10	6	22.4
2	30	32.8	12	5	21.2
3	20	29.4	15	4	20.3
4	15	27.2	**20**	**3**	**20.1**
5	12	25.7	**30**	**2**	**20.1**
6	10	24.6	60	1	21.1

5. Results

5.1. Number of Mel-Frequency Filters vs DCT Order

We first study the effects of the number of mel filters and the number of DCT coefficients by fixing the time-frequency context size to $M = 27$ frames (225 milliseconds), context shift to 18 frames (1/3 overlap), DFT order to $Q = 32$ and the number of Gaussians to 64. The results are shown in Table 2.

Increasing the number of mel-frequency filters improves accuracy as expected, results saturing at $C = 20$ to about 20 % EER and error rates increasing for $C \geq 35$. Regarding the number of DCT coefficients, the best results are obtained either using $D = 2$ or $D = 3$ coefficients whereas the error rates for $D = 1$ and $D = 4$ are systematically higher. The high error rates at the lower right corner of Table 2 are caused by the numerical problems of the GMM classifier: the dimensionality of the features is too high relative to the training sample size and the number of Gaussian components.

One may argue that degradation in accuracy for $D > 3$ is merely because of the increased dimensionality and the associated problems with the statistical model. To gain further insight into the relative importance of the "acoustic" and "modulation" dimensions, we fix the dimensionality to $C \times D = 60$ and study all the parameter combinations. The results are displayed in Table 3.

The best settings are $(C, D) = (20, 3)$ and $(C, D) = (30, 2)$, both yielding the same error rate of EER $= 20.1\%$. For these settings, $C \gg D$, which suggests that the acoustic frequency resolution is more crucial than the modulation frequency resolution. On the other hand, increasing the number of mel filters to $C = 60$ shows an increase in the error rate which indicates that the joint frequency representation is useful, though the improvement is not much.

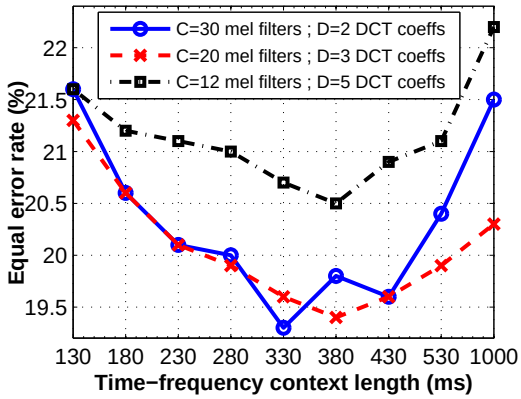

Figure 3: Effect of the time-frequency context size.

5.2. Context Length

Another interesting issue is the effect of the time-frequency context length M. For an increased M, the resolution of the modulation spectrum can be increased, and it is reasonable to hypothesize that we would need more DCT coefficients to model the increased details of the modulation spectra. We select the settings $(C, D) = (30, 2)$, $(C, D) = (20, 3)$ and $(C, D) = (12, 5)$ and vary the context length M. In all cases, we fixed the FFT order to $Q = 256$ and adjusted the context shifts to obtain approximately same number of training vectors for all context lengths. In this way, having dimensionalities and training set sizes equal, any differences in the accuracies can be attributed to the context length and not to the statistical model.

The result is shown in Fig. 3. The settings with more resolution on the acoustic frequency give better results for all context lengths which is consistent with the previous experiment. For all the three settings, the error curves are convex and show optimum context sizes either at 330 ms ($M = 41$ frames) or 380 ms ($M = 47$ frames). From the settings $(C, D) = (30, 2)$ and $(C, D) = (20, 3)$, the latter one with more DCT coefficients gives better accuracy at very long contexts as hypothesized.

5.3. Further Optimizations

Next, we fix $(C, D) = (30, 2)$ and $M = 41$ (330 milliseconds) and further fine-tune the system by using voice activity detector (VAD) [7], and increasing the number of GMM components to 128. Adding the VAD reduces the error rate from EER=20.1 % to EER=18.1 % and increasing the model size to 128 reduces it further to EER=17.4 %. In our previous study [6], we reported an error rate of EER=25.1 % on the same data set by using the full modulation spectrogram with a long-term averaging classifier and T-norm score normalization. We conclude that the accuracy of the modulation spectrogram classifier has been significantly improved by a combination of dimension reduction, better classifier and VAD.

5.4. Comparison with MFCCs

Finally, we compare the proposed feature with the conventional MFCCs. Our MFCC GMM-UBM system [15] first computes 12 MFCCs from a 27-channel mel-frequency filterbank. The MFCC trajectories are then smoothed with RASTA filtering, followed by delta and double delta feature computation. The last two steps are voice activity detection and utterance-level

Table 4: Comparison of MFCC and modulation spectrogram based features and their fusion (EER %) for different test segment durations.

Test duration (s)	MFCC	Mod.spec.	Fusion
0–20	10.5	18.6	10.5
20–30	8.5	17.6	8.4
30–40	7.6	16.6	7.3
40–60	7.7	15.8	7.3

mean and variance normalization. The same GMM-UBM classifier setup is used for both feature sets.

The accuracies across different test segment lengths are shown in Table 4. The results for the different test lengths were obtained by extracting the corresponding scores from the trial list and the fusion result is obtained by a linear combination of the log likelihood ratio scores. The weights of the fusion were optimized using the FoCal toolkit[1] which minimizes a logistic regression objective function.

Overall, the accuracy of the MFCC-based classifier is higher as expected. For short test segments, the fusion is not successful. For longer test segments, there is a slight improvement, which is an expected result. The modulation spectrum measures low-frequency information which is likely to be more subject to degradation for very short test segments. Nevertheless, the fusion gain is only minor. It would be interesting to study further the accuracy by using significantly longer training and test segments, such as those found in the 3- and 8-conversation tasks of the NIST SRE 2006 corpus.

6. Conclusions

We have presented a dimension reduction method for the modulation spectrogram feature and studied its performance in the single-channel speaker verification task. Mel-frequency filtering and DCT were used for reducing the number of acoustic and modulation spectrum coefficients, respectively. The best results were obtained using 20 to 30 mel filters, 2 or 3 DCT coefficients and a context length of 330 to 380 milliseconds. This context length is significantly longer than the typical time span of delta and double-delta features, and similar to those used in speech recognition [3]. The best overall accuracy on the NIST 2001 set was EER=17.4 % which is significantly better than our previous result of EER=25.1 %.

The conventional MFCC feature outperformed the proposed feature in terms of accuracy. A slight improvement was obtained when combining these two features with linear score fusion. The modulation spectrum based feature cannot be yet recommended for applications. Further experiments with longer training and test data are required to confirm whether the contextual features would benefit from larger training set sizes.

Both in this paper and in [7], we used the signal-independent DCT in feature extraction, mostly due to its energy compaction property. However, it might not be the best method for speaker verification. Ideally, we should select or emphasize those modulation spectrogram components which discriminate speakers and are robust against channel mismatch and noise. It is not clear which these frequencies would be. In [11], modulation filtering of the mel-filter outputs indicated that modulation frequencies between 1-4 Hz would be most important, whereas frequencies below 0.125 Hz and above 8 Hz would be harmful for recognition. In that study, however, same filtering operation was applied to all mel-frequency subbands, which does not take advantage of the joint information between the acoustic and modulation frequencies. The question of which regions in the joint frequency representation are relevant, remains open.

[1] `http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm`

7. References

[1] H. Hermansky, "Should recognizers have ears?" *Speech Communication*, vol. 25, no. 1-3, pp. 3–27, Aug. 1998.

[2] H. Yang, S. Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Communication*, pp. 35–50, May 2000.

[3] N. Morgan, Q. Zhu, A. Stolcke, K. Snmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, . Cetin, H. Bourland, and M. Athineos, "Pushing the envelope - aside," *IEEE Signal Processing Magazine*, pp. 81–88, Sept. 2005.

[4] W. Campbell, J. Campbell, D. Reynolds, D. Jones, and T. Leek, "Phonetic speaker recognition with support vector machines," in *Proc. Neural Information Processing Systems (NIPS)*, Dec. 2003, pp. 1377–1384.

[5] B. Ma, D. Zhu, R. Tong, and H. Li, "Speaker cluster based GMM tokenization for speaker recognition," in *Proc. Interspeech 2006*, Pittsburgh, Pennsylvania, USA, September 2006, pp. 505–508.

[6] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Proc. ICASSP'2006*, Toulouse, France, 2006, pp. 665–668.

[7] T. Kinnunen, E. Koh, L. Wang, H. Li, and E. Chng, "Temporal discrete cosine transform: Towards longer term temporal features for speaker verification," in *Proc. 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP 2006)*, Singapore, December 2006, pp. 547–558.

[8] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 668–675, 2003.

[9] S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proc. ICASSP 2007*, vol. 4, Honolulu, Hawaii, USA, April 2007, pp. 605–608.

[10] D. Hardt and K. Fellbaum, "Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification," in *Proc. ICASSP 1997*, Munich, Germany, April 1997, pp. 867–870.

[11] S. Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, November 1998, pp. 3205–3208.

[12] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[13] T. Quatieri, *Discrete-Time Speech Signal Processing - Principles and Practice*. Prentice-Hall, 2002.

[14] S. Vuuren, "Speaker verification in a time-feature space," Ph.D. dissertation, Oregon Graduate Institute of Science and Technology, March 1999.

[15] R. Tong, B. Ma, K.-A. Lee, C. You, D. Zhu, T. Kinnunen, H. Sun, M. Dong, E.-S. Chng, and H. Li, "Fusion of acoustic and tokenization features for speaker recognition," in *Proc. 5th Int. Symp. on Chinese Spoken Language Processing (ISCSLP'2006)*, Singapore, December 2006, pp. 566–577.