JOINT ACOUSTIC-MODULATION FREQUENCY FOR SPEAKER RECOGNITION

Tomi Kinnunen

Institute for Infocomm Research (I^2R)
Media Division / Human Centric Media
Speech and Dialogue Processing Lab
21 Heng Mui Keng Terrace, Singapore 119613
ktomi@i2r.a-star.edu.sg

ABSTRACT

We propose a method for computing joint acoustic-modulation frequency feature for speaker recognition. This feature describes the amplitude modulation spectrum of each subband, and results in a single feature vector per utterance. This vector is directly used as the speaker's modulation frequency template, excluding the need for a separate training phase. The effects of analysis parameters and pattern matching are studied using the NIST 2001 corpus. When fusing the proposed feature with the baseline MFCC/GMM system, EER is reduced from 18.2 % to 16.7 %.

1. INTRODUCTION

Speaking is a *dynamic* process, in which the airflow generated by the lungs is continuously modulated by the articulatory movements to produce the acoustic output. Rapid articulatory movements are reflected in the short-term spectral changes, which can be measured by the *delta* features [1, 2]. The delta features are typically appended with the corresponding static features into a long vector at the frame level. The baseline acoustic recognizer in the state-of-the-art text-independent speaker recognition systems is usually a Gaussian mixture model trained on the cepstral vectors [3].

The delta features capture spectral dynamics within very short time intervals ($<$ 50 ms). However, it is widely accepted that linguistically and perceptually relevant *modulation frequencies* of speech are concentrated on the frequency range of 1-20 Hz [4, 5, 6], corresponding to time intervals 50-1000 msec. These low frequency modulation components are not captured by the conventional short-term spectral analysis. As evidenced by many engineering studies, emphasizing the spectral modulations in the range of 1-10 Hz increases speech recognition robustness (e.g [7, 8]). For a recent survey on modulation frequencies, see [6].

In [9], the relevance of different modulation frequencies for speaker recognition was studied by temporally filtering the mel-filter outputs and monitoring the effect to the error rate by excluding certain modulation bands. It was found out that in both matched and mismatched conditions, the modulation frequencies between 1-4 Hz are most relevant, whereas frequencies below 0.125 Hz and above 8 Hz are less relevant.

Unlike in [7, 8, 9], in which band-pass filtering of the modulation spectrum plays an intermediate role in enhancement of other features, we consider the modulation spectrum as a speaker characterizing feature itself. In particular, we propose a method for computing a *joint* frequency representation in which the amplitude modulation spectra of all frequency subbands is included. Following terminology of [6], we call the frequency variable of the original spectrogram the *acoustic* frequency, and the frequency variable of the second transformation along the subband amplitude envelopes the *modulation* frequency. The joint spectral density is a matrix serving as the speaker's modulation frequency template, and it can be visualized as a gray-scale image or a contour plot (Figs. 1 and 2).

2. JOINT ACOUSTIC-MODULATION FREQUENCY

The computation of the joint acoustic-modulation frequency spectrum is carried out in three phases (see Fig. 1). In the first stage, spectrogram is computed using conventional short-term Fourier analysis. This is followed by another short-term Fourier analysis along the DFT output amplitude envelopes. Finally, these short-term modulation spectra are time-averaged. The implementation details and motivation for each step are given in the following subsections.

2.1. Time-Frequency Representation

The speech signal $s(n)$ is analyzed in short overlapping frames [10]. Each frame is preemphasized and multiplied by a Hamming window before the DFT computation [11]. The modulus of each DFT bin is computed. We denote the resulting time-frequency representation as $S(n, \omega)$, where n denotes the discrete time variable (frame number) and ω denotes the discrete frequency variable (DFT bin).

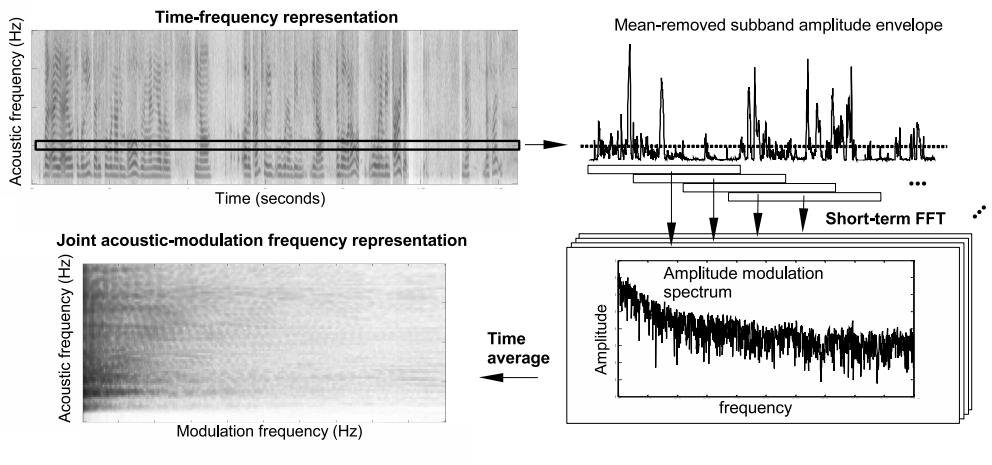

Fig. 1. Computation of joint acoustic-modulation frequency representation.

Selection of the frame length and frame rate for the computation $S(n, \omega)$ is crucial. A long frame implies increased acoustic frequency resolution, with the cost of decreased time resolution. Stated in another way, rapid spectral changes (i.e. high modulation frequencies) cannot be detected with a long window. By using a short window, the bandwidth of the modulation spectrum is increased. This bandwidth also depends on the type of the window function. For the Hamming window, the theoretical bandwidth[1] of the modulation spectrum (B) is given by $B = 2f_s/L$, where f_s denotes the sampling rate of the signal and L is the length of the window in samples [12].

Given the knowledge that the window function length limits the bandwidth of the modulation spectrum, we can determine the frame rate automatically. The frame rate f_r of the spectrogram $S(n, \omega)$ is the sampling rate of the modulation spectrum. In order to avoid aliasing, sampling rate must be selected higher than twice the bandwidth, i.e. $f_r > 2B$. Thus, when given the maximum modulation frequency of interest $\eta_{\max}$ (Hz), we set the frame rate to $f_r = \lceil 2\eta_{\max} \rceil$, and use a window of length $L = \lceil 2f_s/\eta_{\max} \rceil$ samples. For instance, for $\eta_{\max} = 20$ Hz and $f_s = 8$ kHz, we have $f_r = 40$ frames/sec and $L = 800$ samples (or 100 msec).

2.2. Joint Acoustic-Modulation Frequency

After obtaining the subband amplitude envelope $S(n, \omega), n = 0, 1, 2, \ldots$ for a fixed $\omega = \omega_k$, the mean of the amplitude envelope is subtracted [4]. This removes the zero frequency or DC component of the modulation spectrum which represents static information.

For the mean-removed subband envelope, we perform spectral analysis in overlapping Hamming-windowed frames. The window length of this analysis (M) specifies the modulation frequency resolution, and it is considered the other control parameter of the method in addition to the maximum modulation frequency $\eta_{\max}$. In this study, we fix the analysis frame shift of the modulation spectrum to $(2/3)M$. Figure 2 shows an example of the effect of control parameters to the bandwidth and the resolution of the modulation spectrum.

In order to make the resulting modulation spectrum less dependent on the text, the modulation spectra are time-averaged, a method that was used to reduce within-speaker variance already in the 1970s [13].

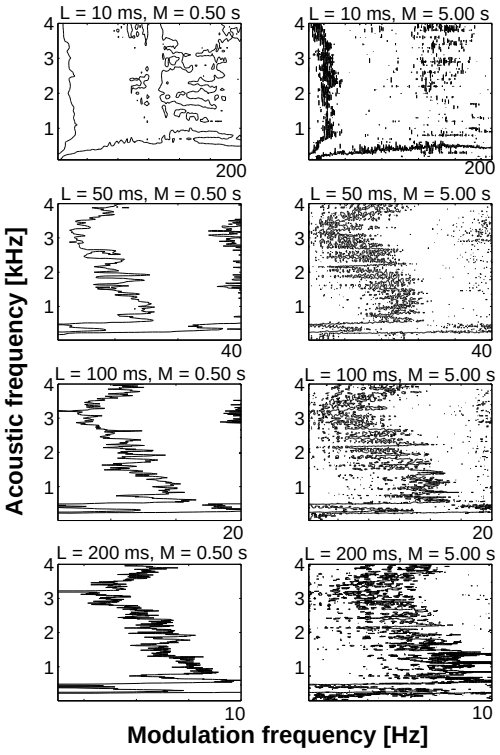

Fig. 2. Effect of analysis parameters to modulation spectrum.

[1] defined as the location of the first zero of the window function magnitude response.

3. PATTERN MATCHING

Since the modulation frequency matrix has fixed dimensions, we simply reshape it into a vector. Given the reference modulation vector $\boldsymbol{m}_r$ and the test utterance modulation vector $\boldsymbol{m}_t$, we consider four simple similarity measures: Euclidean distance, correlation coefficient, cosine of the angle between the vectors, and symmetric Kullback-Leibler distance. For the Kullback-Leibler distance [14], the vector elements are divided by the sum of all elements so that the vector represents a probability mass function. Note that the Euclidean metric depends on the absolute scale, whereas the other metrics are scale-invariant, and thus expected to be more robust.

In speaker recognition, normalization of match scores relative to other models increases robustness [3, 15]. In this study, we adopt the *test normalization* method ("Tnorm") [15], in which the unknown utterance X is first matched against a set I of pseudoimpostor models. The mean $\mu_I(X)$ and standard deviation $\sigma_I(X)$ of the pseudoimpostor scores are then obtained. Finally, the normalized score s' is obtained from the raw score s as $s' = (s - \mu_I(X))/\sigma_I(X)$.

4. EXPERIMENTS

For the experiments, we use the cellular data of the one-speaker detection task from the NIST 2001 corpus [16]. The corpus is provided with a development set and an evaluation set, which do not have speaker overlap. We use the male data (38 speakers) of the development set for parameter tuning. The whole development set (60 speakers) is used as the pseudoimpostor set for Tnorm. The evaluation set consists of 9350 male trials (850 genuine + 8500 impostor) and 13068 female trials (1188 genuine + 11880 impostor). There is about 2 minutes of training data per speaker, and the length of the test segments varies from a few seconds up to one minute.

4.1. Parameter Tuning

The parameters were varied as follows: $\eta_{\max} \in \{5,10,15,20,50,100\}$ (Hz) ; $L \in \{0.3, 0.5, 1.0, 3.0, 5.0, 10.0\}$ (sec). For each parameter combination, the equal error rate (EER) was computed on the tuning set. The best parameter combinations and the corresponding EERs are given in Table 1. The average EER and standard deviations over the 36 parameter combinations are also given, and Fig. 3 shows the complete error surface of the cosine measure as an example.

The Euclidean distance performs worse compared with the other measures. This is not surprising because it is affected by the scale of the spectrum, which is likely to vary. The other three measures are comparable with each other. More importantly, the best performances are obtained by parameters close to each other: bandwidth of 20 Hz and window size of 0.3-0.5 seconds. Based on these observations, we fix the parameters for the evaluation set as $(\eta_{\max}, L) = (20\text{Hz}, 0.3\text{sec})$. The dimension of the modulation frequency vector with this setting is 3200.

4.2. Results on the Evaluation Set

The DET curves for the unnormalized and the Tnorm scores are shown in Fig. 4. In all cases, normalization improves accuracy. Similar to the tuning set, the Euclidean measure performs the worst and the other methods are close to each other in accuracy. The Tnormalized Kullback-Leibler distance gives the smallest error rate.

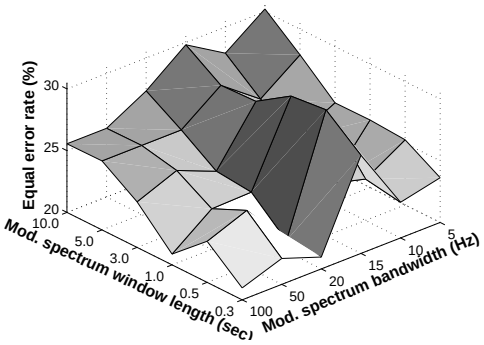

Fig. 3. Error surface for the cosine measure.

Table 1. Results for the tuning set.

	Eucl.	KL dist.	Corr.	Cos.
Best				
$(\eta_{\max}, L)$	(50,0.5)	(20,0.3)	(20,0.3)	(20,0.5)
EER (%)	36.9	20.6	18.9	20.9
Average				
EER (%)	39.5±2.1	24.5±3.0	24.8±2.9	25.1±2.7

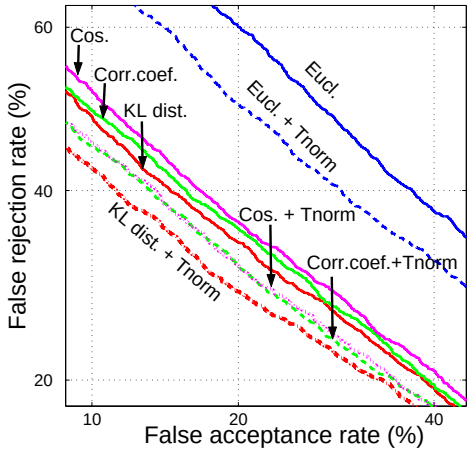

Fig. 4. Comparison of the similarity measures.

4.3. Combining Cepstral and Modulation Features

Finally, we combine the modulation frequencies with conventional MFCC features modeled with adapted GMM [3]. Twelve MFCC coefficients are appended with their delta and double-delta features, followed by global mean subtraction and variance normalization. A 256-component background model is trained using all the speech files from the development set of the NIST 2001 corpus. Target models are trained by adapting the mean vectors from the background model using a relevance factor of 15 (see details in [3]). We use the average log-likelihood ratio as the match score. The fast scoring algorithm using top 10 mixture components as explained in [3] is used.

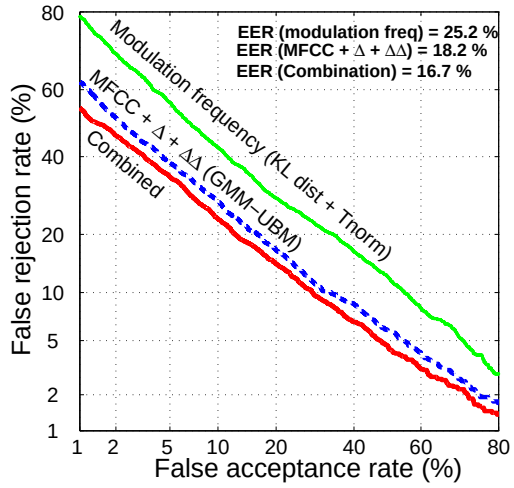

Fig. 5. Combination of modulation and cepstral features.

Based on the previous experiment, we consider only the test normalized Kullback-Leibler distance for the combination with the GMM. The sign of the Kullback-Leibler distance was flipped, followed by constant adding to make the two scores compatible. After experimenting with the weighted sum and product combination rules, the best combination rule was found to be $s = D^{0.2} \times L^{0.8}$, where D denotes the Kullback-Leibler similarity and L the average log likelihood ratio. The resulting DET curve and equal error rates are given in Fig. 5. The modulation features give a slight improvement consistently over different decision thresholds.

The best result for the NIST 2001 benchmark test is less than 10 % EER [17], which is significantly smaller compared with our result (16.7 %). One explanation might be that, due to time limitations, we restricted the GMM size to 256 components - typically the GMM size is 512-2048. Thus, in order to get closer to the state-of-the-art accuracy, the GMM configuration should be first revised. Also, we expect further improvements to the modulation spectrum by using discriminative classification instead of simple time averaging.

5. CONCLUSIONS

In this study, we introduced a modulation frequency feature for text-independent speaker recognition and reported preliminary results on the NIST 2001 corpus. The best result (25.2 % EER) on the evaluation set was obtained by computing the modulation spectrum over a 300 msec window and using a bandwidth to 20 Hz. A slight improvement was also obtained when combining modulation spectrum with the conventional static and dynamic cepstra. Our future plans include exploring discriminative classification, as well as addressing the question of data limitation - how much speech is needed in order the modulation spectrum to saturate.

References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[2] F.K. Soong and A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.

[3] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Proc.*, vol. 10, no. 1, pp. 19–41, 2000.

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

[5] S. Greenberg and T. Arai, "The relation between speech intelligibility and the complex modulation spectrum," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, 2001, pp. 473–476.

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

[7] H. Hermansky, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[8] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117–132, 1998.

[9] S.v. Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc. ICSLP 1998*, Sydney, Australia, 1998, pp. 3205–3208.

[10] J.R. Deller Jr., J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, second edition, 2000.

[11] F.J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–84, 1978.

[12] S.v. Vuuren, *Speaker Verification in a Time-Feature Space*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, March 1999.

[13] J.D. Markel, B.T. Oshika, and jr. A.H. Gray, "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 330–337, 1977.

[14] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.

[15] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digit. Signal Proc.*, vol. 10, pp. 42–54, 2000.

[16] "Nist 2001 speaker recognition evaluation documentation," October 2005, `http://www.nist.gov/speech/tests/spk/2001/doc/`.

[17] M. Przybocki and A. Martin, "NIST speaker recognition evaluation chronicles," in *Proc. Speaker Odyssey 2004*, Toledo, Spain, 2004, pp. 15–22.