Spectro-temporal modulation energy based mask for robust speaker identification

Tai-Shih Chi, Ting-Han Lin, and Chung-Chien Hsu
Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan,
tschi@mail.nctu.edu.tw, tinghanlin.1985@gmail.com, hsu.chung.chien@gmail.com

Abstract: Spectro-temporal modulations of speech encode speech structures and speaker characteristics. An algorithm which distinguishes speech from non-speech based on spectro-temporal modulation energies is proposed and evaluated in robust text-independent closed-set speaker identification simulations using the TIMIT and GRID corpora. Simulation results show the proposed method produces much higher speaker identification rates in all signal-to-noise ratio (SNR) conditions than the baseline system using mel-frequency cepstral coefficients. In addition, the proposed method also outperforms the system, which uses auditory-based nonnegative tensor cepstral coefficients [Q. Wu and L. Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," EURASIP J. Audio, Speech, Music Process. **2008**, 578612 (2008)], in low SNR (≤ 10 dB) conditions.

PACS numbers: 43.60.Hj, 43.72.Ar, 43.72.Fx, 43.60.Bf [CG]
Date Received: January 6, 2012 **Date Accepted:** February 29, 2012

1. Introduction

Many algorithms have been developed over the last two decades for speaker recognition applications. Briefly speaking, speaker recognition tackles two problems: speaker identification and speaker verification.[1] Speaker identification is an M-ary testing problem, in which the identity of the speaker from M candidates is to be determined. In contrast, speaker verification is a binary testing problem, in which the system accepts or rejects a speaker's claim of his/her identity. Conventional algorithms for these two problems adopt short-term spectral features such as mel-frequency cepstral coefficients (MFCCs) to build speaker models. These short-term spectral features basically encode short-term smoothed spectral profiles. Systems using these features usually achieve high recognition rates in clean and matched test conditions.[2] However, recognition rates are significantly degraded in mismatched test conditions where speech is corrupted by unknown convolutive or additive noise. Hence, the robustness of speaker recognition has drawn attentions from researchers.[3–6] In the literature on the robustness, speaker identification research usually considers additive noises in a single-session (microphone) setup while speaker verification research usually considers the multi-session variability.

It has been demonstrated that human hearing is very robust against noise in many kinds of recognition tests.[7] Hence, it is intuitive to include properties of human hearing in speech processing systems for potential performance enhancement, such as the speaker identification systems using auditory features.[4,5] Meanwhile, the ideal binary mask (IdBM) was proposed[8] and multiplied to the noisy spectrogram to improve speech intelligibility in various additive background interferences.[9,10] Not surprisingly, estimating the IdBM became a feasible way to boost speech intelligibility perceived by human subjects.[11] However, the IdBM, which is derived from prior known local signal-to-noise ratios (SNRs) of time-frequency (T-F) units of the noisy spectrogram, is not closely related to hearing perception. Humans process sounds in time and frequency domains simultaneously so that people are not affected seriously by pure temporal or pure spectral interferences. Both spectral contents and temporal behaviors of the sound are known prominent identifying features to human hearing. In other words, dynamic spectral shapes, which are presented by spectro-temporal (S-T) modulations of the spectrogram, carry vital information of a sound. These S-T modulations encode structures of the sound[12,13] and are related to speech intelligibility such that objective speech intelligibility measures can be derived by assessing degradations of these modulations.[14] In this paper, we propose a T-F mask based on perception-related spectro-temporal modulation energies (STMEs) to improve speaker identification rates in noisy environments.

Simulations in this paper were conducted in the setup of a single-session text-independent speaker identification system using the TIMIT and GRID (Ref. 15) corpora. Identification rates by applying the proposed mask are compared with rates from systems using conventional MFCC features and the newly developed auditory-based nonnegative tensor cepstral coefficient (ANTCC) features.[5] The rest of this paper is organized as follows. A brief review of the S-T auditory model, which analyzes S-T modulation energies, is given in Sec. 2. The STME based mask is then proposed in Sec. 3. Performance evaluations for speaker identification are demonstrated in Sec. 4 and conclusions are given in Sec. 5 with discussions.

2. Auditory model and features

The S-T analytical auditory model consists of two computational modules.[12] The first module models the function of spectral analysis of the cochlea, and the second module models the function of S-T modulation analysis of the auditory cortex (A1). A brief review of the auditory model is given in this section and much more detailed descriptions of both modules can be found in Ref. 12.

2.1 Cochlear module and auditory cepstral coefficients (ACCs)

This cochlear module models the peripheral auditory system. As described in Ref. 12, the module is firstly comprised of a bank of 128 overlapping asymmetric constant-Q ($Q_{3\ \mathrm{dB}} \approx 4$) bandpass filters, which are evenly distributed over 5.3 octaves to reflect the frequency selectivity of the cochlea. The output of each filter is fed into a non-linear compression stage, a lateral inhibitory network (LIN), and an envelope extractor. The non-linear compression stage models the high-gain saturation of inner hair cells and together with the LIN accounts for the frequency masking effect. In this paper, a simplified linear version which bypasses the non-linear compression stage is used. All tested speech signals were normalized in advance to avoid the non-linear compression of hair cells. Accordingly, outputs of this simplified linear module at various stages can be written as

$$\begin{aligned} y_{ch}[t,f_i] &= s[t] *_t h[t;f_i], \\ y_{lin}[t,f_i] &= \partial_f y_{ch}[t,f_i] = y_{ch}[t,f_i] - y_{ch}[t,f_{i-1}], \\ y_{as}[t,f_i] &= \max(y_{lin}[t,f_i],0) *_t \mu[t;\tau], \end{aligned} \tag{1}$$

where $s[t]$ is the input acoustic signal; $h[t;\, f_i]$ is the impulse response of the ith constant-Q filter with center frequency f_i, $i = 1,\ldots,\ 128$; $*_t$ is the time-domain convolution; $y_{ch}[t,\ f_i]$ is the output of the ith filter, and $y_{lin}[t,\ f_i]$ is the corresponding output of the LIN. The envelope extractor is implemented by a half-wave rectifier followed by a low-pass filter, whose impulse response $\mu[t;\tau] = e^{-t/\tau} \cdot u[t]$ models the current leakage along the neural pathway to the midbrain and $u[t]$ is the unit step function.

The $y_{as}[t,\ f]$, which is the final output of all cochlear filters and is referred to as the auditory spectrogram, represents the T-F envelope/energy distribution of the input sound. Similar to the deviation of MFCCs, auditory cepstral coefficients (ACC[n]) are defined by taking the logarithm and then the discrete cosine transform (DCT) on each frame of the auditory spectrogram as

$$ACC[n;t] = \sum_{i=1}^{128} \log(y_{as}[t,f_i])\cos\left(\pi n\left(i - \frac{1}{2}\right)/128\right). \tag{2}$$

2.2 Cortical module and rate-scale representation

The second module models the S-T selectivity of cortical neurons. In the auditory model, the auditory spectrogram is further analyzed by A1 neurons which are modeled by two-dimensional filters tuned to different S-T modulation parameters, rate and scale.[12] The rate parameter ω (in Hz) depicts how fast/slow the energy of the auditory spectrogram varies along the temporal axis. The scale parameter Ω (in cycle/octave) characterizes how broad/narrow the energy of the auditory spectrogram distributes along the log-frequency (in octave) axis. In addition, these two-dimensional filters also tune to the upward or downward sweeping direction of modulations. This direction selectivity is encoded by the sign of the rate parameter (positive/negative for downward/upward direction). Therefore, the four-dimensional output $r[t, f, \omega, \Omega]$ of this cortical module can be formulated as

$$r[t,f,\omega,\Omega] = y_{as}[t,f] *_{tf} STIR[t,f;\omega,\Omega], \tag{3}$$

where $STIR[t, f; \omega, \Omega]$ is the joint two-dimentional impulse response of the direction selective modulation filter tuned to ω and Ω; and $*_{tf}$ is the two-dimensional convolution in the time and log-frequency domains. More detailed formulations of the $STIR[t, f; \omega, \Omega]$ are available in Ref. 12.

The local energy of the four-dimensional output is then computed as

$$E[t,f,\omega,\Omega] = \Big| r[t,f,\omega,\Omega] + jH[r[t,f,\omega,\Omega]] \Big|, \tag{4}$$

where $H[\cdot]$ is the Hilbert transform along the log-frequency axis. Therefore, for any T-F unit in an auditory spectrogram, the $E[\omega, \Omega; t, f]$, which is referred to as the rate-scale representation, displays local modulation energies pertaining to different combinations of parameters (ω, Ω). As shown in Fig. 1, the left panel demonstrates a sample auditory spectrogram and right panels are corresponding rate-scale representations of the two T-F units indicated by "x" in the auditory spectrogram. Evidently, these two x units have local modulations dominated around {4–16 Hz, 2–4 cycle/octave, upward} and {4 Hz, 1–4 cycle/octave, downward}, respectively. Briefly speaking, the dominant rate represents the local speaking rate and the dominant scale characterizes the local harmonic spacing of the T-F unit.

3. Spectro-temporal modulation energy based T-F mask

We have shown that joint S-T modulations less than 32 Hz and 4 cycle/octave are highly associated with speech structures such that objective speech intelligibility measures can be derived based on degradations of these modulations.[14] In contrast, white noise strongly activates higher rates ($\geq$ 64 Hz) and higher scales (2–8 cycle/octave) as shown in Ref. 16. It indicates that modulations of speech are mostly smoother than modulations of white noise along the time and the log-frequency axes. In this paper, the spectro-temporal modulation parameter space $\Theta = \{(w,\Omega) : 8 \leq |\omega| \leq 32, 1 \leq \Omega \leq 4\}$ is considered *critical* in detecting speech. Scales within the range of $1 \leq \Omega \leq 4$ (cycle/octave) provide high frequency resolution to detect harmonic structures of speech while rates within the range of $8 \leq |\omega| \leq 32$ (Hz) offer high temporal resolution to detect onsets/offsets of speech. Accordingly, a T-F mask $M[t, f]$ is proposed by calculating the *critical* STME to distinguish speech from non-speech unit by unit. The mask is formulated as

$$M[t_i,f_j] = \begin{cases} 1, & \text{if} \displaystyle\sum_{(\omega,\Omega)\in\Theta} E\left[\omega,\Omega;t_i,f_j\right] \geq \delta \max_{\forall t,f}\left(\sum_{(\omega,\Omega)\in\Theta} E[\omega,\Omega;t,f]\right), \\ \eta, & \text{else,} \end{cases} \tag{5}$$

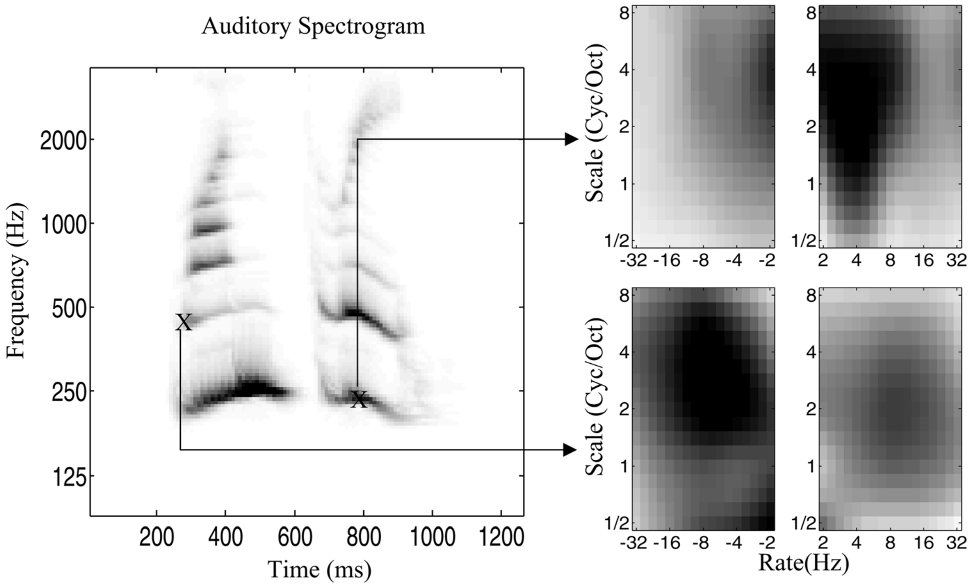

Fig. 1. Rate-scale representation of the cortical module. The left panel shows the auditory spectrogram of a sample utterance "one three" and right panels demonstrate corresponding rate-scale representations of the two T-F units marked by "X." Evidently, the T-F unit around 250 ms has modulations dominated around 4–16 Hz, 2–4 cycle/octave with the upward direction. The other T-F unit around 800 ms has strong modulations around 4 Hz, 1–4 cycle/octave with the downward direction.

where δ acts as a simple threshold on the STME and η as a floor parameter. The values of these two parameters were determined by a pilot simulation as described in Sec. 4. The masked auditory spectrogram is then produced by multiplying the mask $M[t, f]$ to the original noisy auditory spectrogram. Finally, ACCs are extracted from the masked auditory spectrogram for robust speaker identification.

4. Experimental evaluations

In this section, we demonstrate speaker identification rates from our proposed method and compare with rates from systems using MFCCs and ANTCCs. The ANTCCs, which are extracted from the higher-order tensor structure of the cochlear power feature of different speakers, were proposed and shown in Ref. 5 outperforming other two commonly used features, linear predictive cepstral coefficients (LPCCs) and RASTA perceptual linear predictive (RASTA-PLP) coefficients,[17] in speaker identification simulations. In the cited research,[5] the cochlear power feature was very similar to the auditory spectrogram used in our method and noise was removed as minor components in the tensor space. From the functional point of view, the cited research and our proposed method are similar in the way that noisy speech is first represented by an auditory spectrogram then enhanced by a noise reduction mechanism. Hence, we followed the experimental settings in Ref. 5 and compared our identification results with their results.

Two speech corpora, TIMIT and GRID, were used in our simulations. The TIMIT corpus contains clean speech spoken by 630 speakers. As in Ref. 5, 70 speakers were randomly selected from the train folder of TIMIT for evaluations. The first eight utterances and the remaining two utterances per speaker were used as the training and the test sets, respectively. In the GRID corpus, 1000 three-second phrases are spoken by each of the 34 speakers (18 males and 16 females). As in Ref. 5, 50 and 60 sentences per speaker were randomly selected as the training and the test sets. As suggested in Ref. 4, 30 ACCs (excluding the 0th coefficient) per frame were used as our feature vector. The cepstral mean subtraction (CMS) was also adopted for feature normalization.

To focus on evaluating our proposed method, the simple conventional speaker identification algorithm of using Gaussian mixture model (GMM) was considered rather than some sophisticated recognizers. The GMM uses a finite number of multivariate Gaussian functions to approximate the observed probability density function (PDF) of features of a speaker. With its simplicity, fast convergence and good performance, the GMM has become the default reference recognizer in most speaker identification systems. In our system, clean speech was used to build a 32-mixture GMM for each speaker and three types of noises (factory, pink and white) from Noisex-92 database were added in a wide range of SNRs (0, 5, 10, and 15 dB) for test. In addition, the speaker adaptation technique of using a universal background model[18] (UBM) was also considered in our simulations. The UBM, a large GMM constructed from a large amount of speech signals, is usually set with 256–2048 mixtures based on the size of the training data. Lower numbers of mixtures are often used in applications with constrained speech (such as digits or a fixed vocabulary), while 2048 mixtures are more suitable for applications with unconstrained speech (such as the spontaneous speech). In our simulations, a 256-mixture UBM was built using the "sa1" and "sa2" sentences (fixed vocabulary) from all 630 speakers of the TIMIT corpus. The model of each speaker was then derived by adapting parameters of the UBM using his/her training speech signals via the maximum *a posteriori* estimation technique. In our implementations, the relevance factor, which can be viewed as an adaptation coefficient, was fixed to 16 and only the mean of individual GMM speaker model was adapted as suggested in Ref. 18. Detailed formulations and implementations for GMM and UBM can be found in Ref. 18.

To determine the threshold δ and the floor parameter η in Eq. (5), a pilot simulation was conducted using GRID corpus data. Speaker identification rates by our proposed method with various parameter settings in additive pink noise are presented in Table 1. It can be observed that a larger δ (= 0.35) is preferred for lower SNR conditions (5 and 0 dB) but detrimental to higher SNR conditions (15 and 10 dB). Therefore, the two parameters were empirically selected as $\delta = 0.3$ and $\eta = 0.3$ to have better performance under high SNR conditions in our simulations.

Speaker identification rates using MFCCs (the baseline), our proposed ACCs, ACCs with the STME mask (STME_ACC), the STME_ACC with the UBM speaker adaptation technique (STME_ACC_UBM), and the ANTCCs from Ref. 5 are presented in Table 2 for speech samples from the TIMIT and the GRID corpora. Identification rates of using the TIMIT and the GRID corpora are listed on the left and right sides of the slash, respectively. The highest identification rate in each test condition is in boldface. From these test results, one can observe (1) ACCs significantly outperform MFCCs in all test conditions; (2) identification rates are higher using the GRID corpus than using the TIMIT corpus in almost all tests due to the fact that more training data are available for each of the fewer target speakers in the GRID corpus; (3) the STME_ACC produces higher identification rates than ACCs except in the 0 dB GRID corpus condition; and (4) adopting the speaker adaptation technique further improves

Table 1. Speaker identification rates (in %) in additive pink noise with different δ and η parameters using GRID data.

	$\delta = 0.25$				$\delta = 0.3$				$\delta = 0.35$			
	15 dB	10 dB	5 dB	0 dB	15 dB	10 dB	5 dB	0 dB	15 dB	10 dB	5 dB	0 dB
$\eta = 0.5$	90.8	82.2	54.8	25.0	90.9	81.2	59.0	22.8	85.7	74.2	54.6	13.4
$\eta = 0.3$	91.7	82.7	50.2	14.6	90.5	81.5	60.6	19.7	85.9	74.6	57.3	31.9
$\eta = 0.1$	90.2	80.4	30.2	6.86	89.6	83.8	63.7	17.4	85.4	78.3	67.5	41.7
$\eta = 0.01$	88.0	78.0	23.9	3.43	90.1	83.5	66.3	9.41	85.5	81.6	75.8	52.0

Table 2. Speaker identification rates (in %) of 70 speakers (140 sentences) from TIMIT corpus and of 34 speakers (2040 sentences) from GRID corpus.

Noise		MFCC	ACC	STME_ACC	STME_ACC_UBM	ANTCC
type	SNR	TIMIT/GRID	TIMIT/GRID	TIMIT/GRID	TIMIT/GRID	TIMIT/GRID
Factory	0 dB	2.86/5.69	6.43/20.3	11.4/12.3	**16.4/45.7**	2.43/8.82
	5 dB	11.4/24.5	27.9/52.3	57.1/55.8	**62.1/80.6**	12.9/44.6
	10 dB	32.9/51.7	52.9/73.6	75.0/84.5	**85.0/90.3**	49.5/87.8
	15 dB	65.0/73.3	82.1/86.3	85.0/89.7	**92.1**/92.8	78.1/**97.6**
Pink	0 dB	2.86/5.83	4.29/22.5	**12.1**/19.7	10.0/**42.0**	2.43/9.31
	5 dB	5.00/22.6	18.6/47.9	45.0/60.6	**56.4/80.9**	13.8/45.1
	10 dB	21.4/48.0	42.1/70.8	64.3/81.5	**78.6/89.4**	51.0/87.8
	15 dB	46.4/69.1	73.6/83.9	80.0/90.5	**88.6**/92.8	78.6/**95.6**
White	0 dB	2.86/8.92	11.4/32.1	12.1/19.7	**15.0/46.2**	2.86/10.3
	5 dB	10.0/27.5	20.7/40.2	37.9/45.8	**52.1/74.6**	3.81/38.2
	10 dB	18.6/38.6	34.3/55.5	51.4/72.7	**68.6/84.6**	29.5/69.6
	15 dB	36.4/52.2	51.4/72.3	67.9/83.3	**83.6**/88.9	64.3/**95.6**
clean		**100/100**	98.6/95.7	98.6/94.1	98.6/95.0	97.6/**100**

performance of STME_ACC. Comparing with results in Ref. 5, the STME_ACC significantly outperforms ANTCCs in all test conditions using TIMIT data and in low SNR conditions ($\leq$5 dB) using GRID data. However, ANTCCs performs the best in 15 dB conditions using GRID data. A possible reason is that a more distinctive tensor space was constructed for ANTCCs when dealing with speech samples with higher SNRs from the GRID corpus, which contains more training utterances for each of the fewer target speakers.

5. Conclusion and discussions

Human hearing is not only sensitive to spectral contents of a sound, but also to dynamic changes of its frequency components. Such joint S-T properties are very important to speech perception. For instance, the pitch track is strongly emphasized when analyzing the prosody of speech and is considered a vital cue to speech emotions. In this paper, we propose an algorithm to estimate each T-F unit of the noisy auditory spectrogram as either speech or non-speech by assessing its STME. Unlike conventional frame-by-frame energy-based voice activity detection algorithms, which distinguish speech from non-speech by the total energy per frame, our proposed method detects speech by measuring the energy of *critical* S-T modulations of each T-F unit. A cleaned auditory spectrogram can then be generated by applying the unit-by-unit speech activity mask and is shown to provide robust ACC features to enhance speaker identification rates especially in low SNR test conditions.

In our proposed method, two parameters, the threshold δ and the floor parameter η, are involved in generating the STME based mask. As shown in Table 1, these two parameters roughly bear the tradeoff between the amounts of speech distortion and the residue noise.

Therefore, the optimal selection of the parameter values should be SNR-dependent. Estimating the SNR of the test signal to adopt corresponding optimal parameter values will be pursued in the future. In addition, the proposed T-F mask would inevitably create discontinuities in the cleaned auditory spectrogram along the time and the log-frequency axes. The discontinuities along the log-frequency axis are smoothed by the DCT in Eq. (2) and those along the time axis are not detrimental to performance since temporal trajectories of parameters are seldom considered in speaker identification systems. In contrast, for speech recognition applications, where the temporal trajectory is critical to system performance, a missing-data reconstruction algorithm similar to the one in Ref. 19 is needed to estimate missing T-F units from a pre-trained clean speech model. Incorporating the STME based mask in speech recognition systems is another one of our future interests.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work is supported by the National Science Council, Taiwan under Grant No. NSC 100-2220-E-009-004.

References and links

[1] J. P. Campbell, "Speaker recognition: A tutorial," Proc. IEEE **85**, 1437–1462 (1997).

[2] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," Speech Commun. **17**, 91–108 (1995).

[3] M. Ji, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," IEEE Trans. Audio, Speech, Lang. Process. **15**, 1711–1723 (2007).

[4] Y. Shao and D. L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (2008), pp. 1589–1592.

[5] Q. Wu and L. Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," EURASIP J. Audio, Speech, Music Process. **2008**, 578612 (2008).

[6] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," IEEE Trans. Audio, Speech, Lang. Process. **18**, 90–100 (2010).

[7] R. P. Lippmann, "Speech recognition by machines and humans," Speech Commun. **22**, 1–15 (1997).

[8] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer, Norwell, MA, 2005), pp. 181–197.

[9] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**(6), 4007–4018 (2006).

[10] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**(4), 2336–2347 (2009).

[11] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**(3), 1486–1494 (2009).

[12] T. Chi, P. Ru, and S. A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," J. Acoust. Soc. Am. **118**(2), 887–906 (2005).

[13] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proceedings of the International Conference on Spoken Language Processing* (2007), pp. 506–509.

[14] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun. **41**, 331–348 (2003).

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006).

[16] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu, "Robust emotion recognition by spectro-temporal modulation statistic features," J. Ambient Intell. Human. Comput. **3**(2), 47–60 (2012).

[17] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1994).

[18] D. A. Reyolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process. **10**, 19–41 (2000).

[19] S. Srinivasan and D. L. Wang, "Transforming binary uncertainties for robust speech recognition," IEEE Trans. Audio, Speech Lang. Process. **15**(7), 2130–2140 (2007).