Multiband analysis and synthesis of spectro-temporal modulations of Fourier spectrogram

Tai-Shih Chi[a] **and Chung-Chien Hsu**
Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan
tschi@mail.nctu.edu.tw, hsu.chung.chien@gmail.com

Abstract: The two-dimensional spectro-temporal modulation filtering concept of the auditory model [T. Chi, P. Ru, and S. A. Shamma, J. Acoust. Soc. Am. **118**(2), 887–906 (2005)] is implemented on the Fourier spectrogram. The Fourier magnitude spectrogram is analyzed in terms of its joint spectro-temporal modulations, which embed the temporal dynamics and spectral structures. Instead of iterative projection methods, the overlap-and-add method is adopted to invert modified Fourier spectrograms back to sounds. The proposed framework not only provides a similar spectro-temporal analytical process for sounds as the auditory model but also produces synthesized sounds with better quality in a timely manner, which makes proposed framework feasible to human speech recognition (HSR) applications as well.

PACS numbers: 43.60.Hj, 43.60.Uv, 43.72.Ar [CG]
Date Received: December 21, 2010 **Date Accepted:** February 16, 2011

1. Introduction

Neuro-physiological evidences suggest that neurons of the auditory cortex (A1) tune to different spectro-temporal modulations of input sounds. Two types of neurons are observed with preferences in different spectro-temporal modulation directivities, upward and downward. Based on these findings, an auditory model was then proposed and was documented in detail in Ref. 1. In the model, A1's neurons are considered as spectro-temporal selective filters with either downward or upward directivity. This spectro-temporal analytical auditory model was successfully applied to many applications, such as assessing the speech intelligibility[2] and de-noising a spectrogram in a Wiener-filter fashion.[3] On the other hand, the similar idea of applying two-dimensional (2-D) Gabor filters to the mel-spectrogram was evaluated in automatic speech recognition (ASR) applications. The features extracted from outputs of the (optimal) 2-D Gabor filters are demonstrated robust in speech recognition tasks.[4] In addition, the 2-D Gabor filtering approach was adopted to analyze small patches of the spectrogram for detecting prominent structures, such as the harmonicity, formants, and vertical edges (i.e., onset/offset),[5] or for estimating the pitch[6] then extended to estimate formants by exploiting the temporal change of the pitch.[7]

In general, speech related applications can be divided into two categories for humans or for machines to hear. These two types of applications (HSR: human speech recognition and ASR) emphasize different aspects of speech properties and have different objectives, for instance, to improve the speech quality for HSR or to boost the recognition rate for ASR. As in ASR applications, most studies mentioned in the first paragraph emphasize the analysis of sounds by extracting spectro-temporal features from outputs of an analysis stage or a model. Although time-consuming projection methods were proposed in Ref. 1 and used in Ref. 3 to reconstruct speech from a modified auditory-

[a]Author to whom correspondence should be addressed.

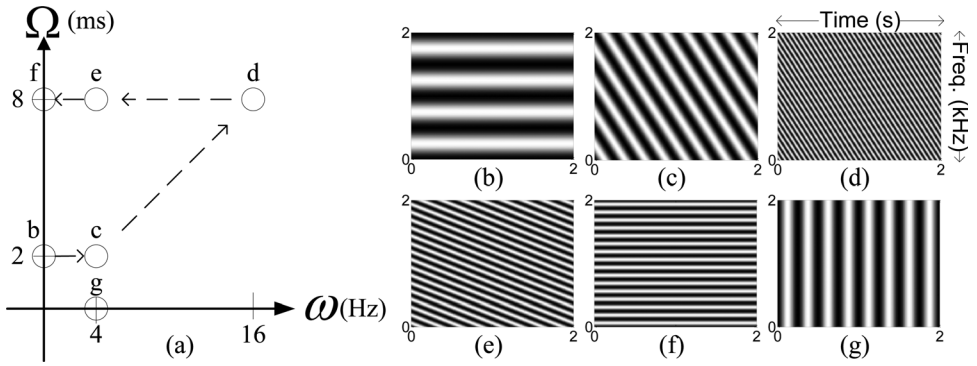

Fig. 1. Time–frequency (T–F) patterns and their corresponding 2-D Fourier transform in the rate-scale (R-S) domain. The rate (ω, in Hz) and the scale (Ω, in ms) are defined as the Fourier domains of the time and the frequency dimensions, respectively. Each panel from (b) to (g) depicts the real part of an analytical T–F pattern, whose imaginary part is the Hilbert transform of the real part along the frequency axis. The corresponding 2-D Fourier transform of each analytical T–F pattern is a single points [points b to g in the panel (a)] in the first quadrant of the ω-Ω space. The values of ω and Ω determine the T–F pattern's densities in the time and the frequency domains, respectively. The ω/Ω ratio determines the slope of the FM sweep of the T–F pattern. Similarly, a single point in the second quadrant of the ω-Ω space maps to an analytical T–F pattern whose real part is with the upward moving directivity rather than the downward moving directivity shown in panels (c), (d), and (e).

spectrogram, distortion occurs during the reconstruction due to nonlinear processes (nonlinear frequency warping from the linear to the logarithm frequency axis and the downsampling along the temporal axis) involved in deriving the auditory-spectrogram. Therefore, these spectro-temporal analysis attempts seem appropriate to ASR applications, but not practical to HSR applications due to reconstruction errors.

In this paper, we propose a spectro-temporal analysis-synthesis framework for the Fourier spectrogram. Along the synthesis path, modified Fourier spectrograms are inverted to sounds conveniently by the overlap-and-add (OLA) method with less distortion, which makes the framework practical to HSR applications. Note, the auditory-spectrogram in Ref. 1 is derived from neither linear processes nor a frame-by-frame scheme such that the OLA method cannot be used for its inversion. To validate our proposed framework, the quality of reconstructed speech is assessed by subjective and objective tests. This paper is organized as follows. In Sec. 2, the proposed spectro-temporal analysis and synthesis framework for Fourier spectrograms is described. We then demonstrate outputs of our analysis process for a sample utterance and address the quality of reconstructed speech from the proposed framework in Sec. 3. The conclusion is given in Sec. 4.

2. Spectro-temporal analysis-synthesis of Fourier spectrogram

The basic idea of the spectro-temporal analysis of the auditory model in Ref. 1 is to treat the auditory-spectrogram as a 2-D image with axes corresponding to time and frequency, then to filter it by using various 2-D modulation bandpass filters, which model the spectro-temporal selectivity of cortical neurons. In this paper, this same idea is applied to conventional Fourier spectrograms.

The magnitude of a Fourier spectrogram of a sound is depicted in the time (T) and the frequency (F) domains. The *rate* (in hertz, as the frequency) and the *scale* (in milliseconds, as the quefrency), which are defined as the Fourier domains of the time and the frequency dimensions, are represented by the parameter ω and Ω, respectively. Any single point in the first quadrant of the ω-Ω space (positive ω), such as the point "c," "d," or "e" in Fig. 1(a), corresponds to a complex time–frequency (T–F) pattern whose real part shows the downward sweeping directivity as in panels (c), (d), or (e) of Fig. 1. On the other hand, any single point in the second quadrant of the ω-Ω space (negative ω) corresponds to a complex T–F pattern with real part showing the

upward directivity (not demonstrated here). Such a complex T–F pattern has the analytical form in which the imaginary part is the Hilbert transform of the real part along the frequency axis. Note, panels (b)–(g) of Fig. 1 only show the real part of corresponding analytical T–F patterns. The panel (b) represents an ideal 500 Hz harmonic complex, whose analytical form maps to a single point [point "b" at $\Omega = 1/500$ s $= 2$ ms in panel (a)] in the ω-Ω domain. The panel (c) shows a downward frequency modulation (FM) complex with joint spectro-temporal modulations of $\Omega = 2$ ms and $\omega = 4$ Hz, whose analytical form has the Fourier transform at the point "c" in the ω-Ω domain. Similarly, the panel (d) depicts a downward FM complex with $\Omega = 8$ ms and $\omega = 16$ Hz modulations. Note, the ω/Ω ratios of point "c" and "d" remain constant such that the FM patterns in panels (c) and (d) have the same slopes but with different densities. For the points "e" and "f," which have the same Ω value as the point "d," their corresponding panels (e) and (f) exhibit the same spectral densities as the panel (d). As the points "b" and "f" correspond to pure harmonic complexes (FM slope $= 0$), the point "g" on the ω axis maps to an analytic signal whose real part shows the FM slope $= \infty$ in panel (g). To sum up, the rate ω and the scale Ω determine the temporal and spectral densities of the T–F patterns, respectively.

In our discrete implementation, two types of joint spectro-temporal modulation filters (STMFs), which are parameterized by ω and Ω for the spectro-temporal decomposition of the Fourier magnitude spectrogram, are generated. The frequency responses of the downward (with subscript "+," positive ω) and the upward (with subscript "–," negative ω) *zero-phase* STMFs can be written as

$$STMF_{+}(\omega,\Omega) = \begin{cases} |\mathcal{F}\{h_{\text{rate}}(t)\} \otimes \mathcal{F}\{h_{\text{scale}}(f)\}|, & 0 \le \omega; \Omega \le \pi \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$STMF_{-}(\omega,\Omega) = \begin{cases} |\mathcal{F}\{h_{\text{rate}}(t)\} \otimes \mathcal{F}\{h_{\text{scale}}(f)\}|, & -\pi \le \omega \le 0; 0 \le \Omega \le \pi \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\mathcal{F}$ is the one-dimensional (1-D) Fourier transform; $\otimes$ is the outer product, and π indicates the half sampling frequency of the discretization process in the time and frequency axes. The h_{rate} and h_{scale} are the 1-D temporal and spectral impulse responses which are derived from gammatone filters centered at frequencies ω_c and Ω_c as

$$\begin{cases} h_{\text{rate}}(t;\omega_c) = t^4 e^{-2\pi BW_{\text{rate}} t} \cos(2\pi\omega_c t) \\ h_{\text{scale}}(f;\Omega_c) = f^4 e^{-2\pi BW_{\text{scale}} f} \cos(2\pi\Omega_c f) \end{cases} \quad (3)$$

where the bandwidths BW_{rate} and BW_{scale} are increased with the center frequencies. The gammatone filters are often used in auditory models to simulate the spectral analysis performed by the basilar membrane.[8]

As shown in Fig. 2, the Fourier magnitude spectrogram of an utterance from a female speaker in the TIMIT corpus is plotted in Fig. 2(a). The speech waveform is first downsampled to 8 k sampling rate and is subject to the short-time Fourier transform (STFT) with a 25-ms hamming window, 5-ms frame shift, and the 800-point fast Fourier transform (FFT) per frame. Therefore, the sampling periods of this 2-D spectrogram in the time and the frequency axes are 5 ms and 10 Hz (8000/800 Hz), which map π to 100 Hz and 50 ms in the rate and the scale domains, respectively. The frequency response of a 2-D noncausal downward STMF, which tunes to $\omega_c = 4$ Hz and $\Omega_c = 2$ ms, is shown in Fig. 2(b). Note, the frequency response of a downward STMF only appears in the first quadrant of the ω-Ω space, which is magnified in Fig. 2(c). Figure 2(d) shows the real part of the corresponding spectro-temporal impulse response (STIR). The output of this modulation filter ($\omega_c = +4$ Hz, $\Omega_c = 2$ ms) to any input Fourier *magnitude* spectrogram $X(t, f)$ can then be written as

$$C(t,f;\omega_c,\Omega_c) = \mathcal{F}_{2D}^{-1}\{\mathcal{F}_{2D}\{X(t,f)\} \cdot STMF_{+}(\omega_c,\Omega_c)\} \quad (4)$$

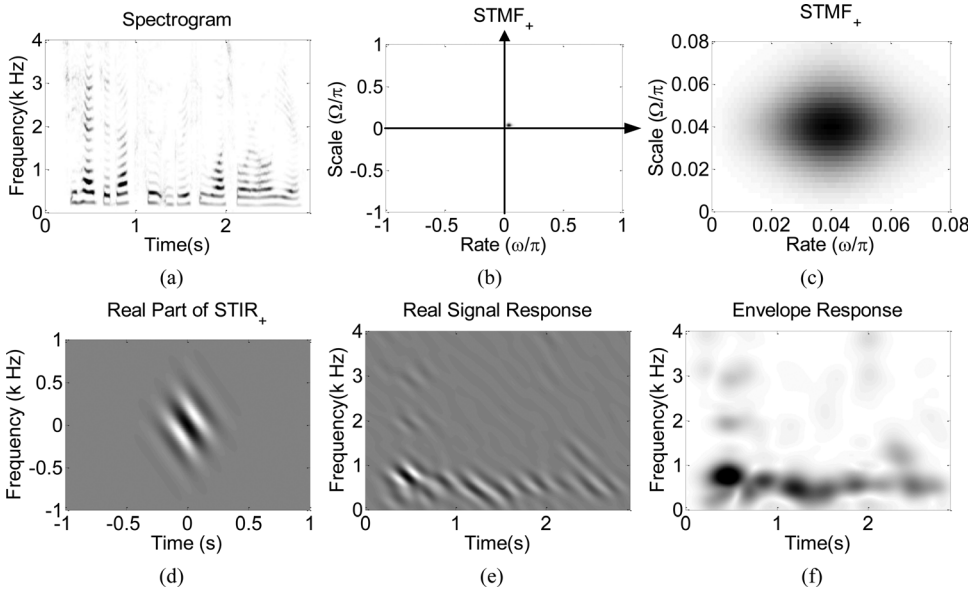

Fig. 2. Input/output and the 2-D frequency/impulse responses of a sample STMF. (a) The Fourier magnitude spectrogram of a sample utterance from the TIMIT corpus as an input to modulation filters; (b) the frequency response of the downward STMF with center frequencies $\omega_c = 4$ Hz and $\Omega_c = 2$ ms; (c) magnified first quadrant of panel (b); (d) the real part of the corresponding analytical spectro-temporal impulse response (STIR); (e) the real part of the output of this sample STMF; and (f) the local energy (envelope) of the output of this sample STMF.

where $\mathcal{F}_{2\mathcal{D}}$ and $\mathcal{F}_{2\mathcal{D}}^{-1}$ denote the 2-D Fourier transform and the inverse 2-D Fourier transform. In addition to the center frequency ω_c and Ω_c, the bandwidth of each 2-D filter can be parameterized as well. In our implementation, this 2-D noncausal zero-phase filterbank is comprised of constant-Q ($Q_{3\text{dB}} \approx 2$ in both ω and Ω dimensions) bandpass filters. Figure 2(e) shows the real part of the output response $C(t, f; \omega_c = +4$ Hz, $\Omega_c = 2$ ms) with the magnitude spectrogram in Fig. 2(a) as the input. Figure 2(f) shows the magnitude response $|(C(t, f; \omega_c = +4$ Hz, $\Omega_c = 2$ ms)|, which reflects the local energy (envelope) of the resolved spectrogram at the spectro-temporal resolution of ω_c and Ω_c.

Since the spectro-temporal analysis is purely a linear filtering operation, the Fourier magnitude spectrogram can be perfectly reconstructed from the four-dimensional (4-D) output representation $C(t, f, \omega, \Omega)$ as long as the 2-D modulation filterbank covers all rate-scale (R-S) components including the DC offset value of the spectrogram. Due to the fact that our 2-D spectro-temporal filters are all with *zero-phase*, the reconstructed spectrogram $X'(t, f)$ can be derived in the synthesis process by the formula shown below:

$$X'(t,f) = \Re\left\{ \mathcal{F}_{2\mathcal{D}}^{-1}\left\{ \frac{\sum_{\omega,\Omega} \mathcal{F}_{2\mathcal{D}}\{C(t,f,\omega,\Omega)\}}{\sum_{\omega,\Omega} STMF_{\pm}(\omega,\Omega)} \right\} \right\} \tag{5}$$

where $\Re\{\bullet\}$ represents the real part of a complex signal. In opposition to iterative projection methods in Ref. 1, the overlap-and-add (OLA) method is utilized to synthesize sounds from the reconstructed magnitude spectrogram $X'(t, f)$. Note, the original phases are preserved for inverting the STFT. The block diagram of proposed spectro-temporal analysis-synthesis framework for Fourier spectrograms is shown in Fig. 3(a).

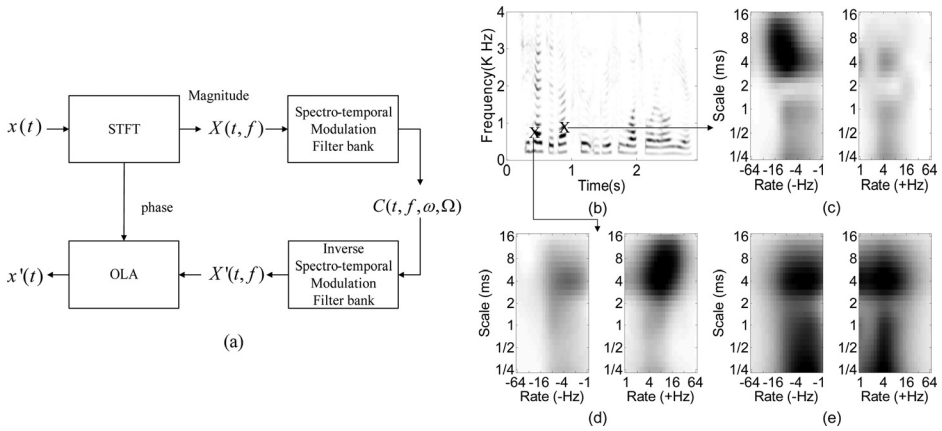

Fig. 3. The block diagram of proposed spectro-temporal analysis-synthesis framework and examples of the R-S plots. (a) The block diagram of proposed framework; (b) a sample magnitude spectrogram as in Fig. 2(a); (c) the R-S plot, which records the local modulation energies at all ω-Ω combinations, of the time–frequency (T–F) unit indicated by x around 850 ms; (d) the R-S plot of the T–F unit indicated by x around 450 ms; and (e) the average R-S plot from all T–F units of the spectrogram in panel (b).

3. Analysis outputs and quality of synthesized speech

An example of the 4-D output $C(t, f, \omega, \Omega)$, which can be used to analyze complex sounds, is given in this section. In our implementation of the analysis process, the rate ω from 1 to 64 Hz and the scale Ω from 0.25 to 16 ms are considered in the spectro-temporal decomposition of Fourier spectrograms of sounds. The sample magnitude spectrogram in Fig. 2(a) is shown again in Fig. 3(b). Figures 3(c) and 3(d) depict the R-S plots, defined as the $|C(\omega, \Omega; t_i, f_j)|$ of a particular (t_i, f_j) unit in a Fourier magnitude spectrogram, of the two T–F units indicated by "x" in Fig. 3(b). As can be observed, for instance, the prominent peak shown in the negative rate in Fig. 3(c) indicates that particular T–F unit possesses strong 16 Hz and 4–8 ms upward spectro-temporal modulations. Figure 3(e) shows the averaged R-S plot over all T–F units in the magnitude spectrogram. The peak around $\omega = 4$ Hz reflects the dominant speaking rate of the female speaker. The strong responses around $\Omega = 1/4$–$1/2$ and 4 ms reveal the dominant frequency spacings between formants and harmonics (pitch $\approx$ 250 Hz) shown in the magnitude spectrogram, respectively. Similar to analyses done in Ref. 1, the 4-D time-frequency-rate-scale output $|C(t, f, \omega, \Omega)|$ can be further integrated over any 2-D for visualizing the characteristic of the sound in the remaining two dimensions. It is worth noting that our proposed spectro-temporal modulation analysis can be degenerated into a pure temporal modulation analysis by only considering modulation filters on the ω axis, such as the point "g" in Fig. 1. This degenerative temporal modulation analysis will produce a three-dimensional (3-D) time–frequency-rate representation. This 3-D representation at any time instant would record the frequency-rate pattern which is referred as the "modulation spectrum" in Ref. 9.

For HSR applications, the speech quality is usually the major concern. Here, the quality of reconstructed speech from our proposed spectro-temporal analysis-synthesis framework for Fourier spectrograms and from the auditory model based analysis-synthesis framework proposed in Ref. 1 is assessed by objective and subjective speech quality measures. In order to address the distortions emerged during the reconstructions, all rates and scales are included to generate the $X'(t, f)$ of clean speech in both frameworks.

The NOIZEUS corpus,[10] which contains 30 phonetically balanced sentences spoken by three male and three female speakers (five sentences per speaker), is used in our evaluations. The mean and the standard deviations of the PESQ (Ref. 11) scores

Table 1. Objective and subjective quality scores of original clean speech and synthesized speech from proposed spectro-temporal analysis-synthesis framework and from the auditory model based (Aud_Model) framework in Ref. 1.

	Objective PESQ score mean/Std.	Subjective MOS score mean/Std.
Aud_model	3.77/0.25	3.35/0.91
Proposed	4.35/0.10	4.33/0.69
Original	4.50/0.00	4.29/0.85

of the original clean speech and the reconstructed speech from both frameworks are given in Table 1, where the "Aud_Model" denotes the auditory model based framework. The PESQ is an intrusive objective speech quality measure, which assesses the degradation between the original speech and the processed speech, with scores ranging from −0.5 to 4.5 (Ref. 12). Therefore, the original unprocessed speech has the 4.5 PESQ score. In subjective listening tests, reconstructed sounds from two frameworks are mixed with original sounds and presented to subjects in a random order through an Audio-Technica headphone. Eight subjects (22–26 yr old with normal hearing) are asked to rate the quality of perceived speech in a five-point scale (1: bad; 2: poor; 3: fair; 4: good; and 5: excellent), which is referred to as the absolute category rating (ACR) scoring. For each sentence, the average of all subjects' ACR scores is called its mean opinion score (MOS).[13] The mean and the standard deviation of the MOS across sentences are also given in Table 1. The reconstructed speech from the proposed framework and the original speech yield almost identical MOS scores within experimental error.

As reported in Ref. 12, the PESQ score for most cases is mapped to a MOS-like score between 1.0 and 4.5, the normal range of MOS values from subjective tests, but may fall below 1.0 in cases with extremely high distortion. Since only reconstructions of clean speech are tested in this study, the PESQ scores represent fair estimates of the speech quality. Both objective and subjective scores in Table 1 indicate the reconstructed speech from proposed framework by the OLA method possesses higher quality than the reconstructed speech from the auditory model based analysis-synthesis framework in Ref. 1. Samples of reconstructed sounds are available at http://perception.cm.nctu.edu.tw/sound-demo.

This spectro-temporal analysis-synthesis framework for Fourier spectrograms can be used in speech applications where reconstructed sounds are needed. For example, we applied the non-negative sparse coding (NNSC) algorithm[14] to the output of each STMF to suppress noises, and then reconstructed a cleaned Fourier magnitude spectrogram for a speech enhancement application.[15] Experimental results demonstrate the reconstructed speech from this sub-modulation-band NNSC algorithm yields higher PESQ scores than enhanced speech by a Wiener-filter, which continuously updates the estimates prior signal-to-noise ratio (SNR) (Ref. 10), in 0–15 dB SNR conditions. In this speech enhancement application, we only modify the sub-modulation-band envelope/magnitude by the NNSC algorithm but leave the phase unchanged. As a result, tonal artifacts are emerged in reconstructed sounds especially in low SNR conditions due to the mismatch between the modified magnitude and the original phase. Therefore, techniques of phase restoration from the magnitude spectrogram, such as the iterative one in Ref. 16, might help to further enhance the quality of reconstructed speech in low SNR conditions and will be investigated in the future development.

4. Conclusion

A joint spectro-temporal analysis-synthesis framework for conventional Fourier spectrograms is proposed in this paper. In opposition to the spectro-temporal analysis done on small patches of the spectrogram,[5,6] where the outputs would be affected by the R-S smearing effects from the 2-D T–F window used in choosing the small patches, our spectro-temporal modulation filterbank decomposes the Fourier magnitude

spectrogram as a whole and reveals the local modulations of each T–F unit. In summary, our proposed framework not only presents a similar joint spectro-temporal analysis process as the auditory model[1] but also reconstructs sounds in a timely manner with less distortion especially when the magnitude spectrogram is modified. Such properties make our proposed framework feasible for both ASR and HSR applications.

Acknowledgments

The authors would like to thank the associate editor and the anonymous reviewers for their comments. This work is supported by the National Science Council, Taiwan, under Grant No. NSC 99-2220-E-009-056.

References and links

[1]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).
[2]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).
[3]N. Mesgarani and S. A. Shamma, "Denoising in the domain of spectrotemporal modulations," EURASIP J. Audio, Speech, Music Process. **2007**(3), 42357 (2007).
[4]B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," Speech Commun. In Press.
[5]T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proceedings of theInternational Conference on Spoken Language Processing* (2007), pp. 506–509.
[6]T. F. Quatieri, "2-D processing of speech with application to pitch estimation," in *Proceedings of the International Conference on Spoken Language Processing* (2002), pp. 1737–1740.
[7]T .T. Wang and T. F. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," IEEE Trans. Audio, Speech, Lang. Process. **18**(1), 171–186 (2010).
[8]R. D. Patterson, M. H. Allerhand, and C. Guiguere, "Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform," J. Acoust. Soc. Am. **98**(4), 1890–1894 (1995).
[9]L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. 761–764 (2004).
[10]P. C. Loizou, *Speech Enhancement: Theory and Practice* (CRC, New York, 2007).
[11]"Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation, ITU-T, Geneva, Switzerland, p. 862 (2001).
[12]J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, Part II-Psychoacoustic model," J. Audio Eng. Soc. **50**(10), 765–778 (2002).
[13]"Methods for subjective determination of transmission quality," ITU-T, Geneva, Switzerland, p. 800 (1996).
[14]P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," J. Mach. Learn. Res. **5**, 1457–1469 (2004).
[15]C.-C. Hsu, T.-H. Lin, and T.-S. Chi, "FFT-based spectro-temporal analysis and synthesis of sounds," Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (In press).
[16]D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust., Speech, Signal Process. **32**(2), 236–243 (1984).