# VOICE ACTIVITY DETECTION BASED ON FREQUENCY MODULATION OF HARMONICS

*Chung-Chien Hsu[1], Tse-En Lin[1], Jian-Hueng Chen[2], and Tai-Shih Chi[1]*

[1]Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan, R.O.C.
[2]Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taiwan, R.O.C.

**ABSTRACT**

In this paper, we propose a voice activity detection (VAD) algorithm based on spectro-temporal modulation structures of input sounds. A multi-resolution spectro-temporal analysis framework is used to inspect prominent speech structures. By comparing with an adaptive threshold, the proposed VAD distinguishes speech from non-speech based on the energy of the frequency modulation of harmonics. Compared with three standard VADs, ITU-T G.729B, ETSI AMR1 and AMR2, our proposed VAD significantly outperforms them in non-stationary noises in terms of the receiver operating characteristic (ROC) curves and the recognition rates from a practical distributed speech recognition (DSR) system.

***Index Terms***—voice activity detection, frequency modulation, spectro-temporal analysis

## 1. INTRODUCTION

Many practical speech processing systems have been deployed in modern world. A VAD module, which distinguishes speech from non-speech in audio streams, is often a crucial component in these systems, such as in telecommunication systems [1][2], in the robust automatic speech recognition system [3] and in the speaker recognition system [4]. However, developing a VAD for noisy environments with low signal-to-noise ratios or for any non-stationary noise is still very challenging.

During past decades, many complex VAD algorithms were proposed. For instance, one algorithm, which assumed speech and noise are Gaussian distributed in the discrete Fourier transform (DFT) domain, detected each speech endpoint using a likelihood ratio test [5]. In addition, noise estimation and adaptation techniques were considered to improve its robustness under non-stationary noise environments but with high computational complexity [6]. Another group of algorithms considered long-term speech information, such as the spectral divergence between speech and non-speech [7], the long-term multiband modulation energy tracking [8] and a novel long-term signal variability measure [9], for robust voice activity detection. Recently, one algorithm was proposed to conquer real-world noise using a support vector machine (SVM) recognizer on amplitude modulation features [10]. In addition, harmonic-related features were used to improve decision accuracy due to their robustness in noisy environments [3][11][12].

Neurophysiological evidences suggest that neurons of the auditory cortex (A1) respond to spectral modulations as well as to temporal modulations of the input sounds. Therefore, A1 neurons can be characterized by spectro-temporal receptive fields (STRFs) and a computational auditory model was proposed accordingly [13]. The multi-dimensional output representation of this spectro-temporal auditory model is highly redundant and was compressed using the tensor decomposition technique for an audio classification application [14]. This concept of spectro-temporal modulation filtering has inspired many engineering approaches, such as using spectro-temporal features for robust speech recognition [15] and speaker recognition [16].

Stemmed from the auditory model, we have proposed a spectro-temporal analysis and synthesis framework for the Fourier spectrogram and extended the conventional Wiener filter to the modulation domain for speech enhancement [17]. We have shown that the proposed spectro-temporal analysis of the Fourier spectrogram can capture prominent acoustic structures, such as pitch, harmonicity, formant, amplitude modulation (AM) and frequency modulation (FM) [17]. The pitch, harmonicity and formants are spectrum-related features which were considered in recently developed VADs [3][11][12]. On the other hand, the AM encodes the long-term variations of the acoustic signal and was included in [8][10][12]. In addition, the spectro-temporal analysis process can capture local FM information of the input acoustic signal. The FM is an important feature for people to recognize speech in noisy environments [18] and, to our best knowledge, has not been considered in any VADs. Therefore, by carefully selecting output features of our spectro-temporal analysis process, we can extract FM structures specifically associated with speech and use those features to build a robust VAD.

In this paper, the frequency modulation energy associated with harmonics is used as a simple and efficient robust speech event measurement. The rest of the paper is organized as follows. Section 2 gives a review of our spectro-temporal analysis process for the Fourier spectrogram and demonstrates modulation contents of speech and noise signals. Then, a voice activity detection algorithm based on valid frequency modulation energy is proposed. Since the proposed VAD is energy-based without any recognizers, it is evaluated against standard VADs just like in [9]. The performance comparisons are demonstrated in section 3. We end in section 4 with some conclusions and discussions.

## 2. PROPOSED METHOD

### 2.1. Mathematical formulation

A1 neurons were modeled as two-dimensional complex filters turned to different spectro-temporal parameters [13]. This concept was applied to the Fourier spectrogram as follows. First the Fourier magnitude spectrogram of observed signal is obtained using short-term Fourier transform (STFT) with 50% overlapping frames. Then the magnitude spectrogram is fed into a bank of complex spectro-temporal modulation filters (STMFs). The frequency responses of the downward (with subscript "+", positive $\omega$) and the

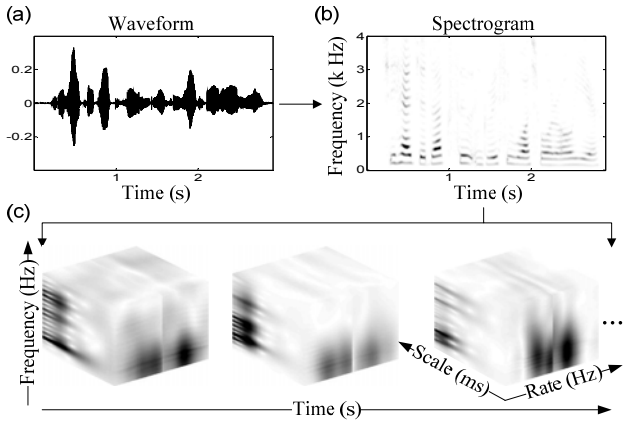


**Fig. 1.** Spectro-temporal analysis of the Fourier spectrogram and the corresponding 4-D output; (a) a sample time waveform; (b) its spectrogram; (c) the 4-D (scale-rate-frequency-time) output.

upward (with subscript "-", negative $\omega$) STMFs can be written as:

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

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

where $\mathcal{F}$ is the 1-D Fourier transform; $\otimes$ is the outer product and $\pi$ indicates the half sampling frequencies of the discrete signal processing along the time and the frequency axes. The rate ($\omega$ in Hz, as frequency) and the scale ($\Omega$ in ms, as quefrency) represent the Fourier domains of the time and the frequency axes, respectively. Note, the complex downward and upward STMFs only locate in the first and second quadrant of the $\omega$-$\Omega$ space, respectively. The $h_{rate}$ and $h_{scale}$ are derived from one-dimensional constant-Q gammatone filter with $Q_{3dB} = 2$. Detailed descriptions can be found in [17].

Therefore, the four-dimensional complex output of the Fourier spectrogram $X(t,f)$ analyzed by the bank of STMFs with different rate-scale parameters can be written as:

$$C(t,f,\omega,\Omega) = \mathcal{F}_{2\mathcal{D}}^{-1}\{\mathcal{F}_{2\mathcal{D}}\{X(t,f)\} \cdot STMF_{\pm}(\omega,\Omega)\} \quad (3)$$

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. Fig. 1 shows the spectro-temporal analysis of the Fourier spectrogram and the corresponding 4-D output. The 4-D local-energy output $|C(t,f,\omega,\Omega)|$ can be further integrated along the frequency axis to produce a local joint spectro-temporal modulation energy profile at any time instant $t_i$ as:

$$E_1(\omega,\Omega;t_i) = \sum_f \left|C(f,\omega,\Omega;t_i)\right| \quad (4)$$

Furthermore, the average joint spectro-temporal modulation energy distribution can be derived by integrating $|C(t,f,\omega,\Omega)|$ along both the time and the frequency axes as:

$$E_2(\omega,\Omega) = \sum_t \sum_f \left|C(t,f,\omega,\Omega)\right| \quad (5)$$

### 2.2. Spectro-temporal analysis

The analysis process can capture spectro-temporal attributes of the input acoustic signal. For speech, the prominent attributes such as

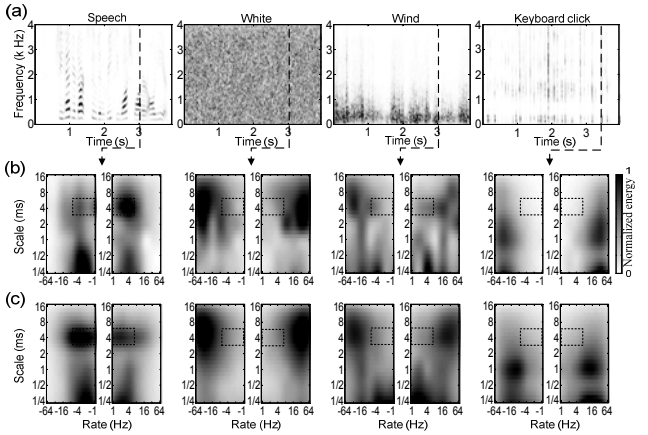


**Fig. 2.** The spectrograms and the corresponding rate-scale patterns of clean speech, white, wind and click noises, respectively. (a) spectrograms; (b) rate-scale patterns at the time instants denoted by the dashed lines; (c) overall averaged rate-scale patterns.

pitch, hamonicity, formant, amplitude modulation, frequency modulation and onset/offset will be resolved dominantly by certain STMFs. Fig. 2(a) shows the Fourier magnitude spectrograms of samples of speech, white, wind and keyboard click noises from left to right respectively. The speech sample was drawn from the TIMIT corpus and the white noise was from the NOISEX-92. The non-stationary wind noise and the keyboard click noise were recorded in real environments. Fig. 2(b) shows their corresponding $E_1(\omega, \Omega; t_i)$ in the rate-scale domain, where the rate ($\omega$) is ranged from 1 to 64 Hz, and the scale ($\Omega$) is from 0.25 to 16 ms, at the time instant denoted by the dashed lines. Furthermore, Fig. 2(c) shows their corresponding $E_2(\omega, \Omega)$ in the rate-scale domain.

The prominent peaks of the rate-scale pattern of speech in Fig. 2(b) reveal that the sample speech is downward moving with a 250 Hz (4 ms) harmonic spacing, a 2000 Hz formant spacing (0.5 ms) and a low temporal modulation (around 4 Hz). As for the white noise, its magnitude spectrogram varies quickly both in the time and the frequency domains such that its rate-scale pattern is strongly activated in the high rate and high scale regions. The non-stationary wind noise has strong energy in low frequency bands and its rate-scale pattern scatters especially in the low rate and low scale regions, which also decode formant information of speech. In other words, the wind noise shares similar formant structures as speech. Unlike the wind noise, the keyboard click noise is an impulse-like noise such that its rate-scale pattern has dominant peaks in the very low scale (due to its frequency content spreading all over the frequency axis) but high rate (due to its transient characteristic) regions. Fig. 2(c) shows the averaged rate-scale patterns across the time axis. From these rate-scale patterns, we can conclude speech and noises distribute differently in the rate-scale domain due to their different acoustic structures. As indicated by these rate-scale patterns, the spectro-temporal modulations resolved by the dashed box region (i.e., harmonics moving downward or upward along the time axis at a low rate) can be treated as crucial structure features for speech/non-speech discrimination. This observation matches the psychoacoustic experiment results that the FM significantly enhances speech reception in noise for human listeners [18].

### 2.3. Frequency modulation energy based VAD

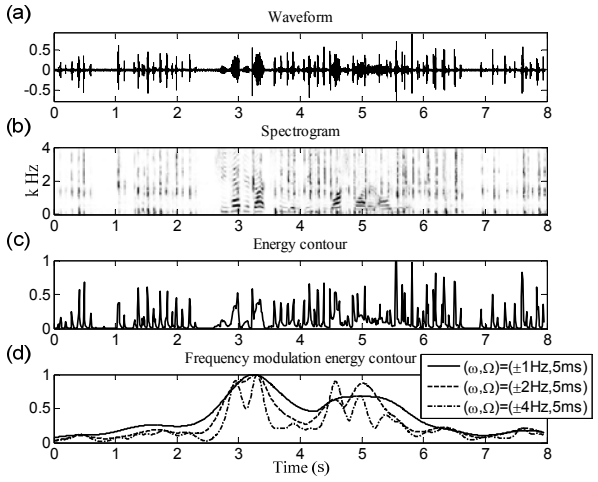


**Fig. 3.** Different energy contours of a speech sample corrupted by 0 dB click noise; (a) noisy speech waveform; (b) spectrogram; (c) energy contour; (d) frequency modulation energy contours from the three parameter settings.

In this section, a VAD algorithm is proposed based on the energy of the frequency modulation of harmonics. To reduce the computational load of the proposed VAD, only a pair of spectro-temporal modulation filters, including one tuned to the upward direction and the other one tuned to the downward direction, are considered in our algorithm. Three settings of $(\omega_c, \Omega_c) \in \{(\pm 1$ Hz, 5 ms), (±2 Hz, 5 ms), (±4 Hz, 5 ms)$\}$ are compared in our evaluations. The bandwidth of the selected $\Omega_c$ = 5 ms constant-Q filter ($Q_{3dB}$ = 2) actually covers 3 to 8 ms, which accounts for the normal harmonic spacing (pitch) range of adult speakers. The Fourier magnitude spectrogram is obtained by STFT using a 20-ms Hanning window with a 10-ms shift. The energy of a certain frequency modulation of harmonics at the time instant $t_i$ is then derived by

$$FME(t_i) = \max_{\omega_c}\{E_1(\omega_c = 4Hz, \Omega; t_i), E_1(\omega_c = -4Hz, \Omega; t_i)\} \quad (6)$$

The frame-by-frame contour $FME(t)$ basically depicts the energies of valid speech frequency modulation structures from our spectro-temporal analysis. Fig. 3(a) presents a sample speech waveform corrupted by keyboard click noises with 0 dB SNR. Fig. 3(b) is the corresponding spectrogram. The regular energy contour and our proposed frequency modulation energy contours of the three settings are depicted in Fig. 3(c) and Fig. 3(d) respectively. All the contours are normalized by their maximum values for display purpose. The speech event is directly determined every 10 ms by comparing the frequency modulation energy with an adaptive threshold.

To obtain the initial threshold, the $FME(t)$ is sorted and divided into 250-ms sections. The section with the lowest value is defined as $FME_N(t)$ and assumed a noise-only section. The section with the highest value is defined as $FME_{S+N}(t)$ and assumed a speech-plus-noise section. The initial threshold is then calculated by

$$\gamma_0 = \rho \cdot \{M[FME_{S+N}(t)] - M[FME_N(t)]\} + M[FME_N(t)] \quad (7)$$

where $M[\cdot]$ is the mean function and $\rho$ is a scaling parameter. The n-th frame of $FME(t)$ of the input signal is compared with the threshold $\gamma_{n-1}$. If the $FME(t = n)$ is larger than the threshold, the current frame is labeled as a speech frame and the threshold is not adjusted. If the current frame is labeled as a non-speech frame, the threshold is updated by integrating $\gamma_{temp}$ and the previous threshold $\gamma_{n-1}$ with a smoothing factor $\alpha$ as in (8).

$$\gamma_n = \begin{cases} \alpha\gamma_{n-1} + (1-\alpha)\gamma_{temp} & \text{for non-speech frame n} \\ \gamma_{n-1} & \text{otherwise} \end{cases} \quad (8)$$

The temporary threshold $\gamma_{temp}$ is calculated from FME of the past 25 candidates of speech frames (stored as $FME_{S+N}(t)$) and non-speech frames (stored as $FME_N(t)$) as in (7).

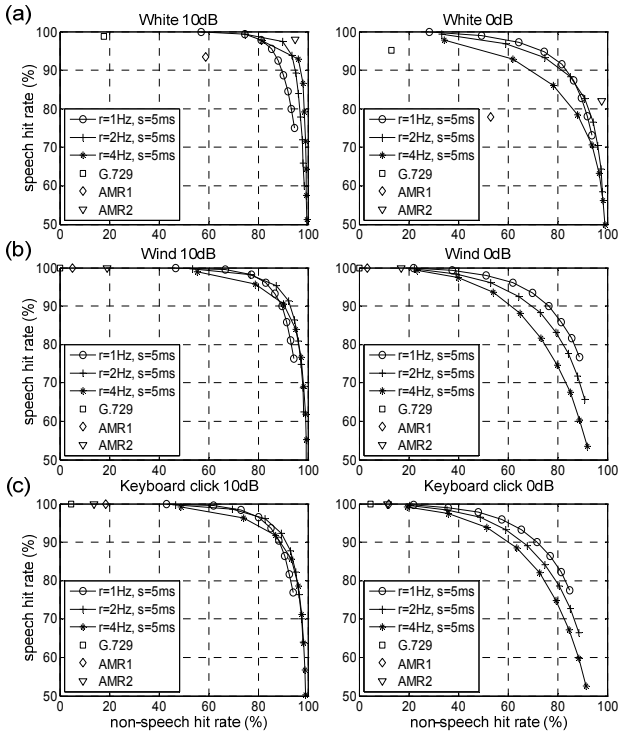


**Fig. 4.** ROC curves of the proposed VAD for two SNR levels and three noise types; (a) white noise; (b) wind noise; (c) computer keyboard click noise.

## 3. EVALUATION AND RESULTS

We conducted a series of experiments to evaluate the proposed VAD. We used the TIMIT test set corpus, which contains 1680 phonetically continuous sentences spoken by 168 speakers (112 male and 56 female speakers) from eight different American dialect regions, in the first part of our evaluations. An average 2-second silence was added to the beginning and the end of each sentence. And as in [19], any less-than-200ms short pause between words was treated as speech. The overall test materials consisted of 38.14% speech and 61.86% non-speech segments, which is close to the active percentage in a typical telephone conversation [20]. In our experiments, noisy signals were generated by adding white, wind and computer keyboard click noises at two SNR levels (10 dB and 0 dB). The desired SNR levels were ensured during speech segments.

The performance of the proposed VAD was assessed using the speech hit rate (H1) and the non-speech hit rate (H0). The speech/non-speech hit rate was defined as the ratio of the number

**Table 1**: DSR recognition rates (%) using different VADs

| | Non-speech interference | Speech-like interference |
|---|---|---|
| | office, street, field_c, field_w | restaurant, bus |
| Hand-labeled | 76.25 | 56.5 |
| G.729 | 52.5 | 23.25 |
| AMR1 | 55 | 30 |
| AMR2 | 60 | 30 |
| CT_VAD | 55 | 26.5 |
| Proposed | 68.75 | 47.5 |

of correctly detected frames to the total number of speech/non-speech frames. The proposed VAD was compared with several standard VADs, the ITU G.729 Annex B (G.729B) [1], the ETSI AMR option 1 and option 2 (AMR1 and AMR2) [2]. In our experiments, the parameter $\rho$ was set as from 0.05 to 0.45 with a 0.05 step and the parameter $\alpha$ was chosen from {0.8, 0.9, 0.98}. The $\alpha$ = 0.98 setting produced the highest VAD accuracy in terms of either H1 or H0. Fig. 4 shows the receiver operating characteristic (ROC) curves of the proposed VAD ( $\alpha$ = 0.98) with respect to $\rho$ for two SNR levels (left: 10 dB; right: 0 dB) and three noise types (top: white noise; middle: wind noise; bottom: click noise). The ROC points of three standard VADs are also given in the figure. A higher $\rho$ results in a higher adaptive threshold such that the H1 is decreased and the H0 is increased. The AMR2 VAD performs the best among the three standards in white noise but neither one of them performs well in non-stationary wind and click noises. Our proposed VAD delivers much higher performance than those standard VADs in wind and click noises and comparable performance as AMR2 in white noise.

Next, the proposed VAD was evaluated in a pilot simulation using a practical DSR system. The on-line DSR system was developed by Chunghwa Telecom Co. for mobile-phone users to automatically search the telephone number of a target institute. The database contains around 60000 telephone numbers of companies and government organizations in northern Taiwan. In our evaluations, we collected 10 10-second recordings in each of the six real environments, including office, street, field_c, field_w, restaurant and bus, through 2G and 3G communication networks. The field_c and field_w refer to the field environments with strong mobile-phone keypad click noise and wind noise. There were 120 test utterances in total and five VADs, including G.729, AMR1, AMR2, CT_VAD (the original VAD in the DSR system) and our proposed VAD, were evaluated. The parameters ($\omega$, $\Omega$, $\alpha$, $\rho$) of the proposed VAD were set as (±1, 5, 0.98, 0.25). The six test environments can be further divided into two categories: environments with non-speech or speech-like inferences. The corresponding average recognition rates (in %) are given in Table 1 with the upper bound from hand-labeled VAD results. Clearly, our proposed VAD outperforms all other VADs in terms of the recognition rate when used in the DSR system.

## 4. CONCLUSION AND DISCUSSIONS

In this paper, we propose a voice activity detection algorithm based on spectro-temporal modulation contents of the input sound. Prominent structures of the input sound can be captured by the spectro-temporal modulation decomposition. In our algorithm, a specific frequency modulation of moving harmonics is assessed and compared with an adaptive threshold to distinguish speech from non-speech. Although harmonic-related features intuitively can only account for vowels, surrounding consonants are still covered due to the low rate filter, which acts as a long term integrator. The ROC curves and recognition rates of the DSR system demonstrate our VAD significantly outperforms standard VADs under non-stationary noise conditions. Because the spectro-temporal modulation analysis works on the Fourier spectrogram, the VAD can be easily integrated into conventional speech processing applications.

Just like in any VADs, the trade-off between the speech hit rate (H1) and the non-speech hit rate (H0) is inevitable in our algorithm. The decision parameter $\rho$ can be set differently based on the application on hand. In the future, we will extend our algorithm to deal with more difficult tasks, such as facing interferences with harmonic structures (for example, sounds from music instruments or animal calls), by carefully re-designing and selecting modulation filters with a more complicated decision mechanism.

## 5. ACKNOWLEDGEMENTS

This research is supported by National Science Council, R.O.C. under Grant NSC 101-2220-E-009-004 and Chunghwa Telecom Co., Ltd.

## 6. REFERENCES

[1] A. Benyassine, E. Shlomot, H. Su, D. Massaloux, C. Lamblin, and J. Petit, "ITU-T recommendation G.729 annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, no. 5, pp. 64–73, 1997.

[2] Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels, ETSI EN 301 708 Rec., ETSI, 1999.

[3] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE J. Select. Top. Signal Process.*, vol. 4, no. 5, pp. 834–844, 2010.

[4] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no.7, pp. 2026–2038, 2011.

[5] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[6] B. Lee and M. Hasegawa-Johnson, "Minimum mean squared error a posteriori estimation of variance vehicular noise," in *Proc. Biennial on DSP for In-Vehicle and Mobile Systems*, 2007.

[7] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, no. 3–4, pp. 271–287, 2004.

[8] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no.6, pp. 2024–2038, 2006.

[9] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 600–613, 2011.

[10] J.-H. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: generalization to novel background classes," in *Proc. ICASSP*, pp. 41–44, 2010.

[11] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Proc. ICASSP*, pp. 4466–4469, 2010.

[12] E. Chuangsuwanich and J. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," in *Proc. INTERSPEECH*, pp. 2645–2648, 2011.

[13] T. Chi, P. Ru, and S. A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

[14] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no.3, pp. 920–930, 2006.

[15] R. M. Stern and N. Norgan, "Hearing is believing: biologically inspired methods for robust automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 34–43, 2012.

[16] H. Lei, B. T. Meyer, and N. Mirghafori, "Spectro-temporal Gabor features for speaker recognition," in *Proc. ICASSP*, pp. 4241–4244, 2012.

[17] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, "Spectro-temporal subband wiener filter for speech enhancement," in *Proc. ICASSP*, pp. 4001–4004, 2012.

[18] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no.7, pp. 2293–2298, 2005.

[19] J. Wu and X.-L. Zhang, "An efficient voice activity detection algorithm by combining statistical model and energy detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, 2011.

[20] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of ITU-T/ETSI voice activity detectors," in *Proc. ICASSP*, pp. 1425–1428, 2001.