IEICE TRANS. INF. & SYST., VOL.E98–D, NO.10 OCTOBER 2015

PAPER

Robust Voice Activity Detection Algorithm Based on Feature of Frequency Modulation of Harmonics and Its DSP Implementation

Chung-Chien HSU[†a)], **Kah-Meng CHEONG**[†], **Tai-Shih CHI**[†], ***Nonmembers***, ***and*** **Yu TSAO**[††], ***Member***

SUMMARY This paper proposes a voice activity detection (VAD) algorithm based on an energy related feature of the frequency modulation of harmonics. A multi-resolution spectro-temporal analysis framework, which was developed to extract texture features of the audio signal from its Fourier spectrogram, is used to extract frequency modulation features of the speech signal. The proposed algorithm labels the voice active segments of the speech signal by comparing the energy related feature of the frequency modulation of harmonics with a threshold. Then, the proposed VAD is implemented on one of Texas Instruments (TI) digital signal processor (DSP) platforms for real-time operation. Simulations conducted on the DSP platform demonstrate the proposed VAD performs significantly better than three standard VADs, ITU-T G.729B, ETSI AMR1 and AMR2, in non-stationary noise in terms of the receiver operating characteristic (ROC) curves and the recognition rates from a practical distributed speech recognition (DSR) system.

key words: *digital signal processor, frequency modulation, spectro-temporal analysis, voice activity detection*

1. Introduction

Speech is the most important bio-signal for human communication. Nowadays, many speech-related applications are developed to facilitate our daily lives. Voice activity detection (VAD), which detects speech segments in an audio stream, is often included in the front-end of speech-related systems, such as in telecommunication systems [1], [2], robust automatic speech recognition system [3] and speaker recognition systems [4], [5]. Therefore, a robust VAD for any noise condition is greatly needed. However, developing a VAD against real-world non-stationary noise is still very challenging for researchers.

Over the past few years, many complicated VAD algorithms were proposed. For instance, one algorithm detected each speech endpoint using a likelihood ratio test by assuming speech and noise signals are Gaussian distributed in the discrete Fourier transform (DFT) domain [6]. In addition, noise estimation and adaptation techniques were considered to improve its robustness under non-stationary noise environments at the cost of high computational loads [7]. Another group of VAD algorithms emphasize long-term speech information, such as considering the spectral divergence between speech and non-speech [8], and a novel long-term signal variability measure [9].

In addition to these engineering approaches, many researchers consider hearing properties in developing VAD algorithms since the human auditory system is capable of detecting competing sound streams in very noisy environments [10]. Speech signals contain rich information in both spectral and temporal domains with their envelopes varying timbrally in frequency and rhythmically in time [11]. The fluctuations of the envelopes across time and frequency axes are referred to as modulations. The importance of spectral modulation [12] and temporal modulation [13] to speech perception was well studied. Temporal modulations reflect dynamic changes of the vocal tract such that they encode rich linguistic information. Psychoacoustic experiments reveal that slow temporal modulations ($\leq$ 16 Hz) of speech are highly related to speech intelligibility [13], [14]. Inspired by these studies, the temporal amplitude modulation of speech has been embraced in VAD algorithms, such as in the long-term multiband modulation energy tracking algorithm [15] and in the support vector machine (SVM) based algorithm with the amplitude modulation spectral (AMS) features [16]. In addition to using the temporal modulations directly, the temporal amplitude modulation transfer function (MTF) was also considered in robust VAD algorithms to restore the temporal envelopes of speech segments in reverberant environments for more accurate detection [17], [18]. Moreover, pitch (fundamental frequency) and harmonics of a voiced sound are perceptually important to human hearing [10], [19], [20]. This property leads to the approach of adopting harmonic-related features in VAD algorithms [3], [21], [22].

Neurophysiological evidence further suggests that neurons of the auditory cortex (A1) respond to joint spectral-temporal modulations of the input sound [23], [24]. Based on the neurophysiological recordings, a computational auditory model has been proposed accordingly [25] and used to derive a noise-robustness representation of speech [26]. On the other hand, psychoacoustic experiments also demonstrate that the joint spectro-temporal modulations are highly related to speech intelligibility [27] and speech comprehension [11]. The concept of using spectro-temporal modulations has since been adopted in many applications, such as speech intelligibility assessment [28], musical instrument identification [29], and robust feature extraction for automatic speech [30], [31] and speaker recognition [32].

Inspired by the auditory model, a spectro-temporal

Manuscript received April 13, 2015.
Manuscript revised June 19, 2015.
Manuscript publicized July 10, 2015.

[†]The authors are with Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan.
[††]The author is with Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan.
a) E-mail: hsuchungchien.cm97g@g2.nctu.edu.tw

DOI: 10.1587/transinf.2015EDP7138

analysis and synthesis framework has been proposed for the Fourier spectrogram and used to extend the conventional Wiener filter to the spectro-temporal modulation domain for speech enhancement [33]. Furthermore, it has been shown that the spectro-temporal analysis of the Fourier spectrogram can capture prominent acoustic "textures", such as pitch, harmonicity, formant, amplitude modulation (AM) and frequency modulation (FM) [34]. The pitch, harmonicity and formants are spectrum-related features have been considered in VAD algorithms [3], [21], [22]. The AM encodes the long-term variations of the envelope of the acoustic signal and was considered as in [15]–[18], [22]. As for the FM, psychoacoustic studies show FM is an important cue for recognizing speech in noisy environments for people with normal hearing [35]. Even hearing-impaired patients can detect dynamic frequency changes [36] so that encoding the FM cue in the cochlear implant can help patients in real listening conditions [37]. To take the advantage of the FM cue, we propose a robust VAD based on the feature of frequency modulation of harmonics in this paper. To our best knowledge, the information of frequency modulation of harmonics has not been considered in any VAD before. The FM cue specifically associated with voice would be extracted from outputs of the spectro-temporal analysis process and used to build a robust VAD.

In today's modern world, portable devices are everywhere and equipped with many speech-related applications. Therefore, the proposed algorithm would be very valuable if it can be implemented in portable devices and function in real time. Often, complex applications are implemented on DSP platforms for a quick demonstration of the prototype systems [38]–[40]. For the same purpose, we implement the proposed VAD algorithm on a DSP platform and evaluate the computational complexity in this paper. The total processing time of the DSP is also demonstrated. The rest of the paper is organized as follows. Section 2 first gives a review of the spectro-temporal analysis process for the Fourier spectrogram and demonstrates modulation contents of speech and noise signals. Then, the VAD algorithm based on local energy of frequency modulation of harmonics is proposed. Section 3 shows details of the hardware implementation on a DSP chip. Simulation results against real-world recorded noise are demonstrated in Sect. 4. Finally, Sect. 5 gives the conclusion and discussion.

2. Proposed Method

2.1 Spectro-Temporal Analysis of Fourier Spectrogram

A1 neurons were modeled as two-dimensional complex filters turned to different spectro-temporal parameters [25]. This concept was applied to the Fourier spectrogram as follows. First, speech can be assumed quasi-stationary and can be analyzed frame by frame using a short analysis window. The short-term Fourier transform (STFT) of the speech signal $x(n)$ can be written as

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)e^{-j2\pi kl/N} \tag{1}$$

where $w(n)$ is the analysis window function; k is the frequency bin index and N is the total number of points in each frame. For speech processing, the Hamming window with 20–40 ms duration is typically used [41]. In the proposed algorithm, the STFT spectrogram $X(n,k)$ is obtained using a 25 ms Hamming window with a 10 ms shift. For any input magnitude spectrogram $|X(n,k)|$, the spectro-temporal analysis stage produces a 4-dimensional multi-resolution output as follows:

$$C_{\pm}(n,k,\omega,\Omega) = |X(n,k)| *_{nk} STIR_{\pm}(n,k;\omega,\Omega) \tag{2}$$

where $STIR_{\pm}(n,k;\omega,\Omega)$ is the spectro-temporal impulse response of the 2-D modulation filter tuned to ω and Ω; $*_{nk}$ denotes two-dimensional convolution along the time and frequency axes. The sign $(\pm)$ represents the sweeping direction of the modulation filters (positive sign refers to the downward direction and negative sign refers to the upward direction). The *rate* parameter ω (in Hz) reflects how fast the local envelope of the magnitude spectrogram varies along the time axis. The *scale* parameter Ω (in ms) reflects how broad the local envelope of the magnitude spectrogram distributed along the frequency axis. They are the Fourier domains of the time and the frequency dimensions, respectively. The property that the convolution in the time domain corresponds to the pointwise multiplication in the Fourier domain is commonly used for 2-D convolution operations to reduce the computational cost using the fast Fourier transform (FFT) [42]. Therefore, Eq. (2) can be rewritten as follows:

$$C_{\pm}(n,k,\omega,\Omega) = \mathcal{F}_{2D}^{-1}\{\mathcal{F}_{2D}\{|X(n,k)|\} \cdot STMF_{\pm}(\omega,\Omega)\} \tag{3}$$

where $\mathcal{F}_{2D}$ and $\mathcal{F}_{2D}^{-1}$ denote the 2-D Fourier transform and the inverse 2-D Fourier transform; $STMF_{\pm}(\omega,\Omega)$ denotes the frequency responses of the 2D spectro-temporal modulation filters. As shown in Fig. 1, the frequency response of a complex downward/upward modulation filter is confined in the first/second quadrant of the $\omega-\Omega$ space. $STMF_{\pm}(\omega,\Omega)$ are derived as follows:

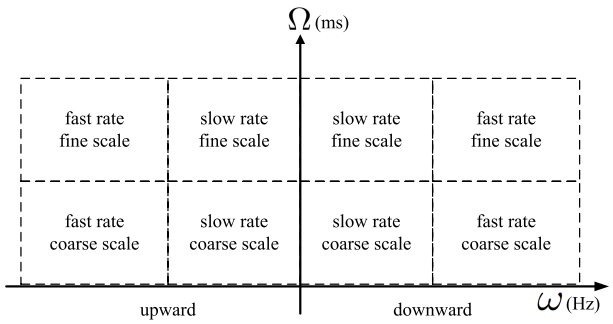

Fig. 1 Rate-Scale $(\omega-\Omega)$ space.

$$STMF_{+}(\omega,\Omega) = \begin{cases} H_{rate}^{min}(\omega;n) \otimes |H_{scale}(\Omega;k)|, & 0 \le \omega \le \pi;\ 0 \le \Omega \le \pi \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$STMF_{-}(\omega,\Omega) = \begin{cases} H_{rate}^{min}(\omega;n) \otimes |H_{scale}(\Omega;k)|, & -\pi \le \omega \le 0;\ 0 \le \Omega \le \pi \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\otimes$ is the outer product; π indicates the half sampling frequency of the sampling process along the time or the frequency axis, respectively. $H_{rate}(\omega;n)$ and $H_{scale}(\Omega;k)$ are the frequency response of the 1-D temporal and spectral modulation filter, respectively. These bandpass modulation filters are formulated as sinusoidal modulated gamma distribution functions [25]:

$$H_{rate}(\omega;n) = \mathcal{F}\{n^4 e^{-2\pi B_{rate} n}\cos(2\pi\omega n)\} \tag{6}$$

$$H_{scale}(\Omega;k) = \mathcal{F}\{k^4 e^{-2\pi B_{scale} k}\cos(2\pi\Omega k)\} \tag{7}$$

where $\mathcal{F}$ is the Fourier transform; the bandwidths B_{rate} and B_{scale} increase with the center frequencies ω and Ω. The impulse response of each $STMF$, which has frequency components in only one quadrant, is a 2D analytic signal such that the corresponding output $C_{\pm}(n,k,\omega,\Omega)$ is full of analytic signals. The $H_{rate}^{min}(\omega;n)$ denotes the *minimum-phase* version of $H_{rate}(\omega;n)$, i.e., $H_{rate}^{min}(\omega;n)$ is the minimum phase system, which has the same magnitude response as $H_{rate}(\omega;n)$. Figure 2 shows the spectro-temporal analysis of a sample magnitude spectrogram $|X(n,k)|$ and the corresponding 4-D output $|C_{\pm}(n,k,\omega,\Omega)|$. The amplitude of an analytic signal was defined as the "local energy" of the signal [43], [44] or can directly represent the local energy [45]. No matter what viewpoints we choose, it is safe to say that the derived $|C_{\pm}(n,k,\omega,\Omega)|$ is highly related to the local energy. In the following contents, the $|C_{\pm}(n,k,\omega,\Omega)|$ is referred to as the "local energy", an abbreviation of "local-energy related feature".

2.2 Rate-Scale Representation

The multi-resolution analysis process can capture the underlying texture of speech. The left panels of Fig. 3 demonstrate the Fourier magnitude spectrograms of samples of speech, white noise, wind noise, and keyboard click noise from top to bottom respectively. The speech sample and the white noise were extracted from the TIMIT corpus [46] and the NOISEX-92 database [47]. The non-stationary wind noise and the keyboard click noise were recorded in real environments. Each 4-D $|C_{\pm}(n,k,\omega,\Omega)|$ can be further integrated along the frequency axis to obtain an averaged rate-scale

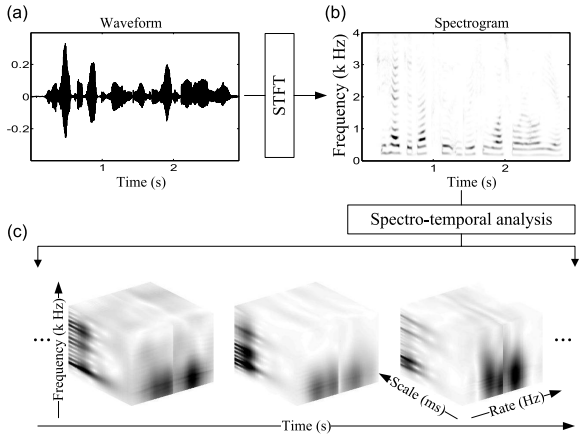

Fig. 2 Spectro-temporal analysis of the Fourier spectrogram and the corresponding 4-D output; (a) a sample time waveform; (b) its spectrogram; (c) the 4-D (rate-scale-frequency-time) output.

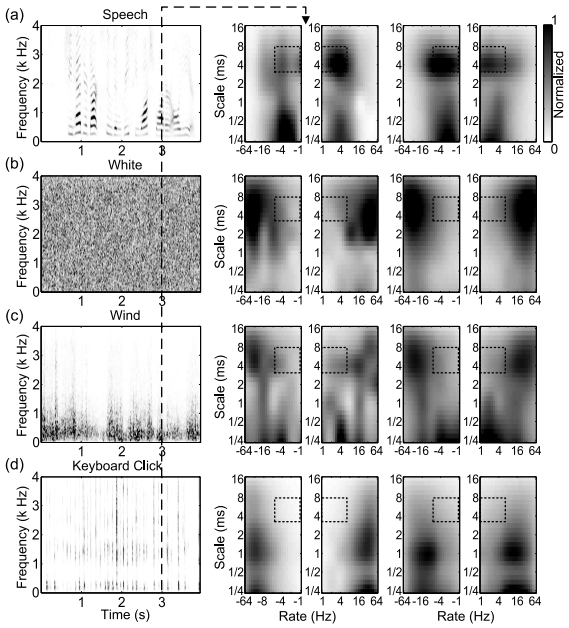

Fig. 3 The spectrograms and the corresponding rate-scale patterns of (a) clean speech; (b) white; (c) wind; (d) click noise, respectively.

representation at the particular time frame as follows:

$$E_{RST}(\omega,\Omega;n_i) = \frac{1}{F}\sum_{j=1}^{F} |C_{\pm}(k_j,\omega,\Omega;n_i)| \tag{8}$$

where F is the total number of frequency bins. The middle two panels of Fig. 3 show the corresponding $E_{RST}(\omega,\Omega;n_i)$ at the time frame n_i (denoted by the dashed line) in the rate-scale domain, where the rate (ω) is ranged from 1 to 64 Hz, and the scale (Ω) is from 0.25 to 16 ms. Note that the positive-rate/negative-rate panel records the output from the downward/upward direction modulation filters. Furthermore, the overall rate-scale representation of the whole signal can be obtained by averaging over both time and frequency axes as follows:

$$E_{RS}(\omega,\Omega) = \frac{1}{L}\frac{1}{F}\sum_{i=1}^{L}\sum_{j=1}^{F}|C_{\pm}(n_i,k_j,\omega,\Omega)| \tag{9}$$

where L is the total number of frames. The right two panels of Fig. 3 show the corresponding $E_{RS}(\omega,\Omega)$ in the rate-scale domain.

The prominent peaks of the rate-scale pattern of speech in the middle two panels of Fig. 3(a) reveal that the envelope of the sample speech signal is downward moving with a harmonic spacing of 250 Hz (4 ms), a formant spacing of 2000 Hz (0.5 ms) and a temporal modulation of 4 Hz. As for the white noise, its magnitude spectrogram varies quickly in both 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 exhibits strong energy in low frequency bands and its rate-scale pattern scatters especially in the low rate and low scale regions. It implies the wind noise shares the similar formant structure as speech but without any harmonic structure. 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. The right two panels of Fig. 3 can be obtained by further collapsing the rate-scale-time patterns over time. From these rate-scale patterns, we can clearly observe speech and noise distribute differently in the rate-scale domain due to their different acoustic textures. 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 critical texture features for speech/non-speech discrimination. This phenomenon provides one possible explanation for the psychoacoustic experiment results that the FM significantly enhances speech reception in noise for human listeners [35].

2.3 FM Local Energy Based VAD

We calculated the *a priori* modulation SNR (SNR^{mod}) over speech segments based on the modulation power in the rate-scale domain as follows:

$$SNR^{mod} = \frac{10}{L_1}\sum_{i=1}^{L_1}\log_{10}\frac{P_{RS}^{speech}(\omega,\Omega;n_i)}{P_{RS}^{noise}(\omega,\Omega;n_i)} \tag{10}$$

where L_1 is the total number of frames in speech segments; P_{RS}^{speech} and P_{RS}^{noise} are the modulation power of speech and noise in each frame, respectively. Similar to Eq. (8), $P_{RS}(\omega,\Omega;n_i)$ is defined as $\frac{1}{F}\sum_{j=1}^{F}|C_{\pm}(n_i,k_j,\omega,\Omega)|^2$. Figure 4 shows the SNR^{mod} calculated from 80 noisy speech signals corrupted by wind noise with 0 dB SNR. Strong responses appear around scale = 4~8 ms. The rate-scale profiles shown in Fig. 3 and Fig. 4 indicate that noise does not uniformly degrade speech in the rate-scale domain. To reduce the computational load of the proposed VAD, only a pair of spectro-temporal modulation filters, one upward and one downward filter, are considered in our algorithm. Three choices of $(\omega_c, \Omega_c) \in \{(\pm 1\text{ Hz}, 8\text{ ms}), (\pm 2\text{ Hz}, 8\text{ ms}), (\pm 4\text{ Hz}, 8\text{ ms})\}$ are compared in our evaluations. The selected $\Omega_c = 8$ ms constant-Q bandpass spectral modulation filter ($Q_{3dB} = 2$) actually covers the scale from 4 to 12 ms. Such filter can resolve the harmonic spacing (pitch) of most adult speakers. Since the moving direction of the harmonics is irrelevant for voice detection, we record the larger local energy out of the two directions as the frequency modulation local energy (FME) of harmonics resolved by (ω_c, Ω_c) at the time frame n_i.

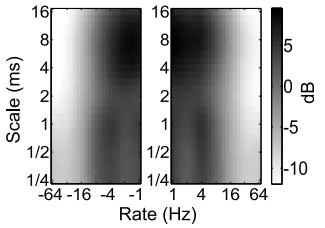

Fig. 4 The *a priori* modulation SNR^{mod} calculated from 80 noisy speech signals corrupted by wind noise at 0 dB SNR.

$$FME(n_i;\omega_c,\Omega_c) = \max_{\omega_c}\{E_{RST}(\omega_{c+},\Omega_c;n_i), E_{RST}(\omega_{c-},\Omega_c;n_i)\} \tag{11}$$

The $FME(n;\omega_c,\Omega_c)$ basically depicted the local energy contour of a valid speech frequency modulation texture of the speech signal. To simplify notation, $FME(n;\omega_c,\Omega_c)$ is abbreviated to $FME(n)$ in following descriptions. Figure 5(a) presents a sample speech waveform corrupted by wind noise at 0 dB SNR. Figure 5(b) is the corresponding spectrogram. The regular energy contour and our frequency modulation local energy contours using three sets of parameters are depicted in Fig. 5(c) and Fig. 5(d), respectively. In addition, Fig. 5(e) shows the pure AM local energy (AME) contours using the three corresponding rate parameters of 1, 2, and 4 Hz. Comparing the non-speech segments (0~2.5 sec and 5.5~8 sec) between Fig. 5(d) and (e), one can easily observe the proposed FM local energy would be much better at distinguishing speech from wind noise than the AM local energy using the approach of thresholding. All the contours are normalized by their own maximum values for display purpose. The voice event is directly determined every 10 ms by comparing the FM local energy with a threshold.

In order to derive the threshold, the $FME(n)$ is first sorted and the resulting sequence is then divided into sections, each of which contains 25 frames. The section containing the smallest value is assumed a noise-only section and the mean of this section is denoted as $\overline{FME}_N$. The section containing the largest value is assumed a noisy section (with both speech and noise) and the mean of this noisy section is denoted as $\overline{FME}_{S+N}$. The threshold γ is then calculated by

$$\gamma = \rho(\overline{FME}_{S+N} - \overline{FME}_N) + \overline{FME}_N \tag{12}$$

where ρ is a scaling parameter. The VAD decisions were

made frame by frame. For the n_i-th frame, it is labeled as a speech frame if $FME(n_i)$ is larger than the threshold. The efficacy of the proposed VAD is demonstrated in Fig. 6. Figure 6(a) shows the waveform of the clean speech sample as in Fig. 5 and its reference VAD labels extracted from the TIMIT word transcriptions. Figure 6(b) shows the corresponding spectrogram and the reference VAD labels. Figure 6(c) depicts the frequency modulation local energy contour of $(\omega_c, \Omega_c) = (\pm 1$ Hz, 8 ms) as in Fig. 5(d) and the threshold calculated by Eq. (12). Figure 6(d) shows the VAD labels from our proposed algorithm.

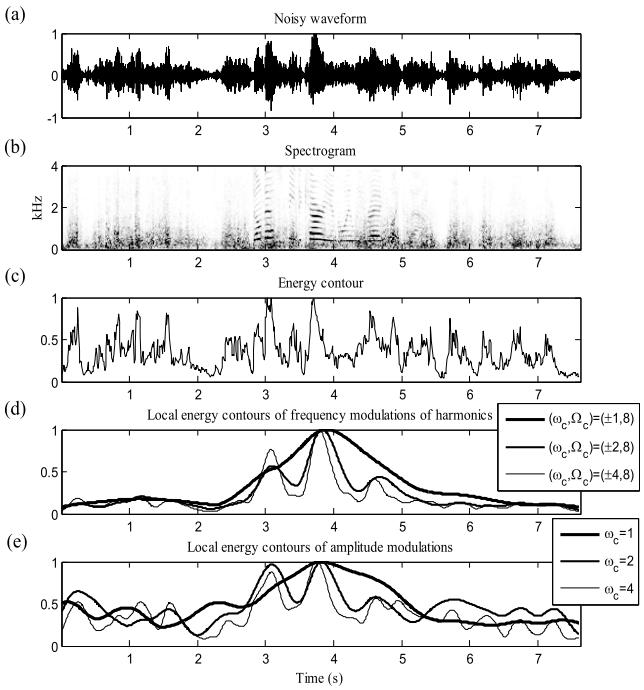

Fig. 5 Different frequency modulation local energy contours of a speech sample corrupted by wind noise at 0 dB SNR. (a) Noisy speech; (b) the corresponding spectrogram; (c) energy contour; (d) frequency modulation local energy contours derived using three sets of parameters; (e) amplitude modulation local energy contours derived using three parameters.

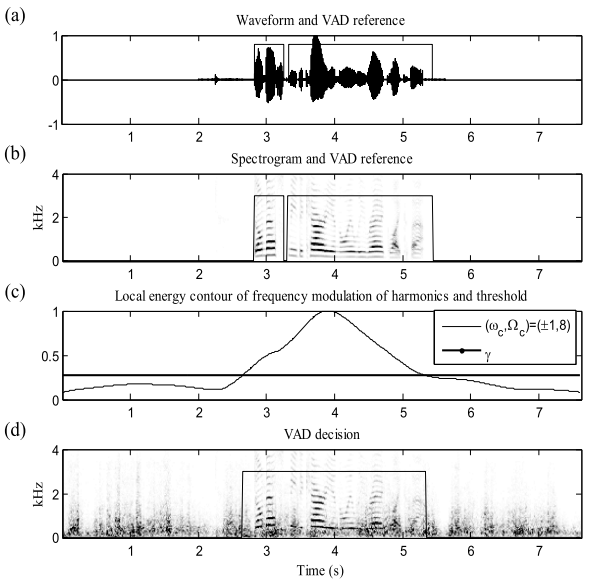

Fig. 6 VAD results from the proposed algorithm for a speech sample corrupted by wind noise at 0 dB SNR. (a) Clean waveform and the reference VAD labels; (b) the clean spectrogram and the reference VAD labels; (c) the frequency modulation local energy contour and the threshold; (d) the noisy spectrogram and the VAD results from the proposed algorithm.

3. Hardware Implementation

It has been shown that DSP chip is more reliable than FPGA [48]. Therefore, for the purpose of prototyping, we implement the proposed VAD algorithm on the TMS320C6713 DSP Starter Kit (DSK), a DSP platform produced by Texas Instruments, for real-time simulation.

The proposed VAD algorithm needs a lot of FFT computations, each of which requires a great amount of data (depending on the FFT length). From our experience, the time of transmitting data by CPU is about 4 times longer than the time of executing the algorithm. Therefore, we need direct memory access (DMA) modules for transmitting data to reduce the overall processing time. Since the TMS320C6713 is a floating-point DSP and provides 16 independent enhanced DMA channels, we implemented the proposed VAD algorithm on TMS320C6713 DSK for convenience.

3.1 Flow Chart

The procedures of the proposed VAD algorithm are summarized in the following list and the corresponding flow chart is shown in Fig. 7.

1) The input acoustic signal was received sample by sample from the analog to digital converter (ADC) with the sampling frequency of 8 kHz. The data was then buffered for the frame-based FFT operation. In our algorithm, each frame (25 ms) contains 200 samples and consecutive frames are 80 samples (10 ms) apart.
2) The 512-point radix-2 complex FFT was applied to

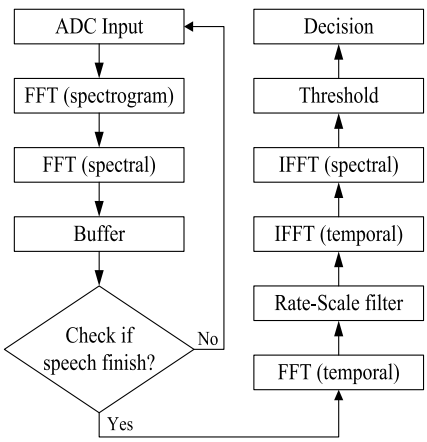

Fig. 7 Flow chart of the proposed VAD algorithm.

Table 1 Complexity per frame (per sample) of the proposed algorithm.

Stage	MUL	ADD
Fourier Spectrogram	9,730	14,081
Rate-Scale Analysis	17,131	24,377
Rate-Scale Synthesis	31,623	47,435
Threshold and Decision	517	514
Overall	59,001 (295)	86,407 (432)

convert the time-domain signal to the Fourier spectrogram frame by frame.

3) The 512-point radix-2 complex FFT was again applied to convert each frame of the Fourier spectrogram into the scale domain. The outcome was called scalogram and was stored for cross-frame temporal processing into the rate domain later.
4) Step 1) to 3) was repeated until all samples of the input speech signal were acquired.
5) The 1024-point radix-4 complex FFT was applied to the scalogram for cross-frame temporal processing. Then the outcome multiplied by the frequency responses of a pair of rate-scale filters. The frequency responses of the rate-scale filters were pre-stored in the memory.
6) 2 inverse FFTs (IFFTs) were applied to convert results in the rate-scale domain back to the time-frequency domain.
7) Calculate the threshold based on Eq. (12) and make VAD decisions frame by frame.

3.2 Complexity and Processing Time

Table 1 provides the computation complexity per frame of the proposed VAD algorithm in terms of the number of real multiplications (MULs) and real additions (ADDs). The complexity of each FFT is equivalent to 2Nlog2N MULs and 3Nlog2N ADDs, where N is the FFT length. The total complexity per frame is 59001 MULs and 86407 ADDs. Since a frame contains 200 samples, the total complexity per sample is 295 MULs and 432 ADDs. Table 2 shows the overall processing time of our algorithm for an 8-second speech signal. The processing cycle was obtained using the profile clock count provided by TI code composer studio (CCS) and the processing time was obtained by dividing the processing cycle with the clock rate of the DSP chip, which is 225 MHz in our case. Several procedures were adopted for the implementation on the DSP platform to reduce the processing time.

1) The radix-4 and radix-2 complex FFT computation was optimized by following TI's optimization procedures.
2) The square root function was replaced by a lookup table.
3) The bubble sorting method was modified for computing the threshold.
4) The linear assembly coding was adopted.

Since the Fourier spectrogram of the input signal is computed during the data acquisition period, its processing time is not included in the overall processing time. As noted in Table 2, the time for transmitting data is 734 ms, which can be totally diminished by using EDMA. Therefore, the overall processing time was reduced to about 204 ms plus overhead of EDMA for an 8-second input speech signal. Finally, the VAD results of real-time emulations on the DSP platform were identical to VAD results of MATLAB simulations on the PC.

Table 2 Overall CPU processing cycle (processing time) of the proposed algorithm.

Stage	Cycle for executing algorithm	Cycle for transmitting data
Fourier Spectrogram	N/A	N/A
Rate-Scale Analysis	11,950,500 (54 ms)	50,886,000 (226 ms)
Rate-Scale Synthesis	31,150,000 (138 ms)	114,210,000 (508 ms)
Threshold and Decision	2,698,600 (12 ms)	N/A
Overall	45,799,100 (204 ms)	165,096,000 (734 ms)

4. Experiment

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. A 2-second silence was added to the beginning and the end of each sentence. The overall test materials consisted of about 38% speech and 62% non-speech segments. In our experiments, noisy signals were generated by adding white noise, wind noise, and computer keyboard click noise at four SNR levels (10 dB, 5 dB, 0 dB, and −5 dB). The desired SNR levels were ensured within speech segments [49].

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 is defined as the ratio of the number of correctly detected speech/non-speech frames to the total number of speech/non-speech frames. Note, a perfect VAD provides 100% H1 and 100% H0. The proposed FM local energy based VAD was compared with three standard VADs, the VADs of ITU-T G.729 Annex B (ITU-T G.729B) [1], ETSI adaptive multi-rate codec option 1 and option 2 (ETSI AMR1 and AMR2) [2], and the AM local energy based VAD shown in Fig. 5(e). The G.729B was developed for fixed telephone and multimedia communication systems and the AMR codec was developed for 3G mobile communication systems. Each noisy speech signal was normalized to −26 dBov [50] before fed into the three standard VADs. In our experiments, the threshold for our proposed VAD was derived by Eq. (12) and the threshold of the AM local energy based VAD was derived by the same equation but using AME instead. In addition, the scaling parameter ρ was set as from 0.05 to 0.60 with a step of 0.05. Figure 8 shows the receiver operating characteristic (ROC) curves of all compared VADs with respect to ρ for four SNR levels and three noise types ((a) white noise; (b) wind noise; and (c) click noise). The ROC curve gives a full description

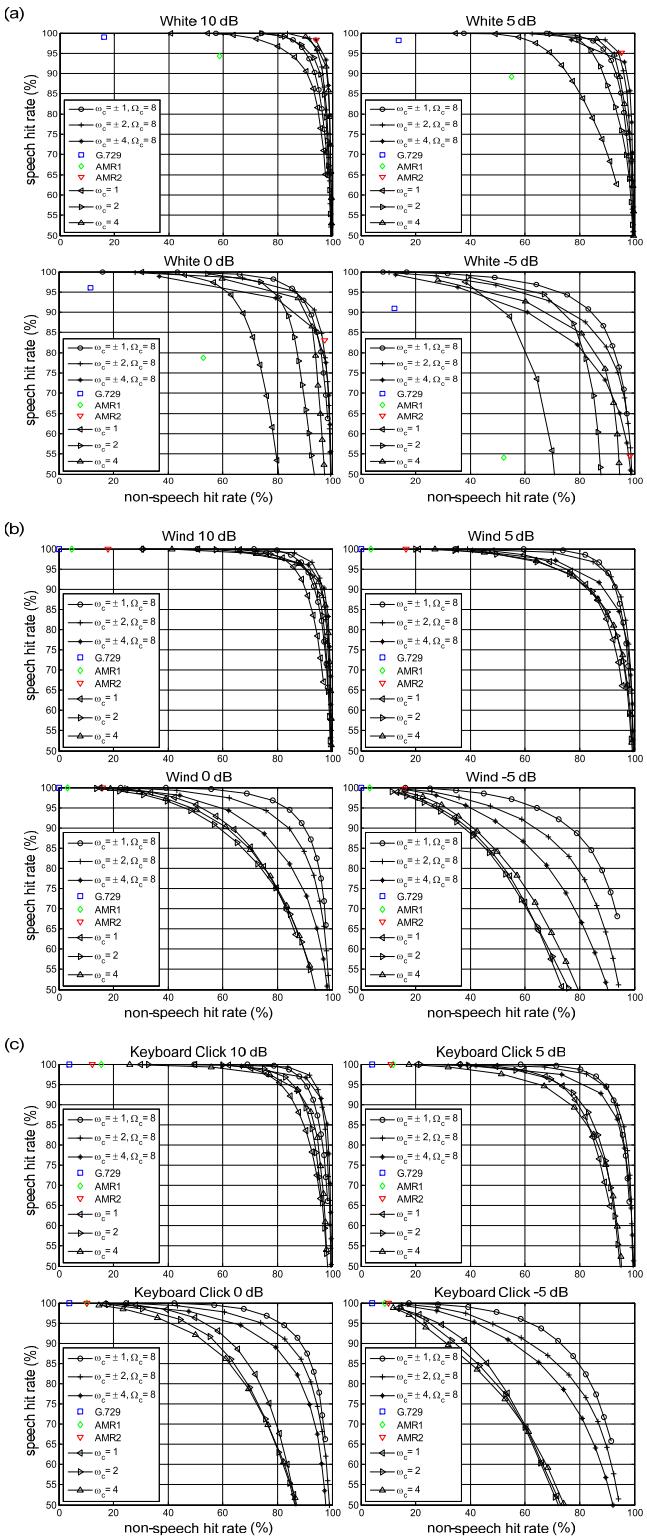

Fig. 8 ROC curves of the proposed VAD for four SNR levels and three noise types; (a) white noise; (b) wind noise; (c) computer keyboard click noise.

of the relationship between H1 and H0. A higher ρ results in a higher threshold such that the H1 is decreased and the H0 is increased. The ROC points of three standard VADs are

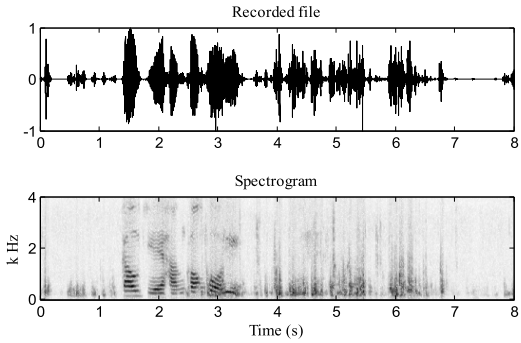

Fig. 9 One noisy speech sample recorded from outdoors with strong wind noise and the corresponding spectrogram.

also given in the figures. The ETSI AMR2 is a very sophisticated VAD, which detects speech based on many factors, including the energies of 16 frequency subbands, the energy of background noise, the channel SNR, the frame SNR, and the long-term SNR [40]. From the simulation results, we can observe that the AMR2 VAD performs the best among the three standard VADs in white noise but neither one of them performs well in non-stationary wind noise and click noise. Our proposed VAD delivers much higher performance than the standard VADs in wind noise and click noise and comparable performance as AMR2 in white noise. Comparing the proposed FME based VAD with the AME based VAD, one can observe that both types of VADs produce comparable results under high SNR (10 dB) conditions. Furthermore, the FME based VAD outperforms the AME based VAD when SNR decreases. The performance gap becomes larger and larger with lower and lower SNR. To sum up, the FME based VAD with parameters (ω_c, Ω_c)=($\pm$1 Hz, 8 ms) produces the best results in low SNR conditions.

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. to automatically search the telephone number of a target institute for mobile-phone users. The database contains around 60000 telephone numbers of companies and government organizations in northern Taiwan. In our evaluations, we collected 10 8-second recordings in each of the six real environments through each of the 2G and 3G communication networks. The test environments include office, street, restaurant, bus, outdoors with strong wind noise, and outdoors with strong mobile-phone keypad click noise. Figure 9 shows a sample of recorded signal from outdoors with strong wind noise and the corresponding spectrogram. There were 120 test utterances in total for each of the five VADs, including G.729B, AMR1, AMR2, CT_VAD (the original VAD used in the DSR system) and the proposed VAD. The parameters (ω_c, Ω_c, ρ) of the proposed VAD were set as ($\pm$1, 8, 0.25). The corresponding average recognition rates (in %) are given in Table 3 with the upper bound from man-labeled VAD results. Clearly, our proposed VAD outperforms all other VADs in terms of the average recognition rate when used in the DSR system. For speech signals recorded in a quiet en-

Table 3 DSR recognition rate (%) for using different VADs of the proposed algorithm.

Method	Accuracy (%)
G.729B	44.55
AMR1	48.18
AMR2	51.82
CT_VAD	47.27
Proposed	64.55
Man-labeled	70.91

vironment such as the office, all VADs produce good results. The main advantage of the proposed VAD is its robustness against non-stationary noise.

5. Conclusion and Discussion

In this paper, we propose a voice activity detection algorithm in the spectro-temporal modulation analysis and synthesis framework [34]. Prominent textures of the input sound can be captured by the spectro-temporal modulation decomposition. In the proposed algorithm, the local energy of the specific frequency modulation of moving harmonics is assessed and compared with a threshold to distinguish speech from non-speech. Conventional VADs usually consider the overall energy, while the proposed VAD only considers the energy attributed to textures of speech. Although harmonic-related textures intuitively can only be attributed to vowels, surrounding consonants can still be covered by the proposed VAD due to the low-rate filter, which acts as a long-term integrator. One side effect of using the long-term integrator is the inevitable inclusion of the short silence between speech segments as shown by the detection error around 3.2 seconds in Fig. 6. To avoid this false alarm error, information of the transients (such as onsets and offsets) embedded in the high rate region in the modulation domain needs to be considered. For instance, as shown in Fig. 5(d), the FM local energy contour resolved by the 4 Hz rate modulation filter depicts voiced segments (i.e., segments with harmonic structures) more accurately than the contour resolved by the 1 Hz rate modulation filter. This accuracy will become higher when using a higher rate modulation filter. However, if we only adopt the high rate information without including the long-term integrator, the unvoiced sections in speech segments would be misclassified as non-speech due to their lack of harmonic structures. Therefore, depending on applications, the high rate and low rate information need to be carefully balanced to address the trade-off between the false alarm errors from short silences and miss errors from short unvoiced sections. In this study, we assume the cost of the miss error is much higher than the cost of the false alarm error such that the high rate information is not considered in our system.

The ROC curves and recognition rates of the DSR system demonstrate our VAD significantly outperforms the three standard VADs under non-stationary noise conditions. The successful implementation on the DSP platform also demonstrates the algorithm can be used in many kinds of speech applications, especially for batch processing applications. Because the spectro-temporal modulation analysis works on the Fourier spectrogram, the proposed VAD can be easily integrated into conventional speech processing applications. In this work, we only use a simple decision method and only consider a pre-selected pair of spectro-temporal modulation filters to validate the idea of using functional energy instead of energy to identify speech segments. In the future, we will develop a more complicated decision rule and an adaptive mechanism of selecting more effective modulation filters based on the user's current background noise to further improve the performance. Another potential direction is to regularize the proposed VAD similar to the approach in [51] to build a more robust system for variations of SNRs and noise conditions . In addition, we will implement the proposed VAD algorithm on the TMS320C6416 fixed-point DSP.

Acknowledgments

The authors would like to thank Jian-Hueng Chen and Yi-Cheng Chen at Chunghwa Telecom Co., Ltd., Taiwan for their help in conducting DSR experiments. The authors would also like to thank the associate editor and the anonymous reviewers for their valuable comments. This research is supported by the National Science Council, Taiwan under Grant No. NSC 102-2220-E-009-049.

References

[1] A. Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin, and J.P. 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.9, pp.64–73, 1997.
[2] ETSI, "Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels," ETSI EN 301 708 Recommendation, 1999.
[3] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," IEEE J. Sel. Topics Signal Process., vol.4, no.5, pp.834–844, 2010.
[4] M.W. Mak and H.B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," Comput. Speech Lang., vol.28, no.1, pp.295–313, 2014.
[5] 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, Language Process., vol.19, no.7, pp.2026–2038, 2011.
[6] 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, Jan. 1999.
[7] B. Lee and M. Hasegawa-Johnson, "Minimum mean squared error a posteriori estimation of high variance vehicular noise," Proc. Biennial on DSP for In-Vehicle and Mobile Systems, 2007.
[8] J. Ramírez, J.C. Segura, C. Benítez, 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.
[9] P.K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," IEEE Trans. Audio, Speech, Language Process., vol.19, no.3, pp.600–613, 2011.
[10] A.S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, 1990.

[11] T.M. Elliott and F.E. Theunissen, "The modulation transfer function for speech intelligibility," PLoS Comput. Biol., vol.5, no.3, p.e1000302, 2009.
[12] M. ter Keurs, J.M. Festen, and R. Plomp, "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am., vol.91, no.5, pp.2872–2880, 1992.
[13] R. Drullman, J.M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am., vol.95, no.2, pp.1053–1064, 1994.
[14] R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," Science, vol.13, pp.303–304, 1995.
[15] G. Evangelopoulos and P. Maragos, "Multiband modulation energy tracking for noisy speech detection," IEEE Trans. Audio, Speech, Language Process., vol.14, no.6, pp.2024–2038, 2006.
[16] J. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," Proc. IEEE ICASSP, pp.41–44, 2010.
[17] M. Unoki, X. Lu, R. Petrick, S. Morita, M. Akagi, and R. Hoffmann, "Voice activity detection in MTF-based power envelope restoration," Proc. INTERSPEECH, pp.2609–2612, 2011.
[18] S. Morita, M. Unoki, X. Lu, and M. Akagi, "Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments," Proc. IEEE ISCSLP, pp.108–112, 2014.
[19] S. Norman-Haignere, N. Kanwisher, and J.H. McDermott, "Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex," J. Neurosci., vol.33, no.50, pp.19451–19469, 2013.
[20] S. Shamma and D. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," J. Acoust. Soc. Am., vol.107, no.5, pp.2631–2644, 2000.
[21] L.N. Tan, B. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," Proc. IEEE ICASSP, pp.4466–4469, 2010.
[22] E. Chuangsuwanich and J.R. Glass, "Robust voice activity detector for real world applications using harmonicity and modulation frequency," Proc. INTERSPEECH, pp.2645–2648, 2011.
[23] D.A. Depireux, J.Z. Simon, D.J. Klein, and S.A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol., vol.85, no.3, pp.1220–1234, 2001.
[24] N. Mesgarani and E.F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," Nature, vol.485, pp.233–236, 2012.
[25] T. Chi, P. Ru, and S.A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," J. Acoust. Soc. Am., vol.118, no.2, pp.887–906, 2005.
[26] N. Mesgarani, S.V. David, J.B. Fritz, and S.A. Shamma, "Mechanisms of noise robust representation of speech in primary auditory cortex," Proc. Natl. Acad. Sci. U.S.A., vol.111, no.18, pp.6792–6797, 2014.
[27] T. Chi, Y. Gao, M.C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am., vol.106, no.5, pp.2719–2732, 1999.
[28] M. Elhilali, T. Chi, and S.A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Commun., vol.41, no.2-3, pp.331–348, 2003.
[29] K. Patil, D. Pressnitzer, S. Shamma, and M. Elhilali, "Music in our ears: The biological bases of musical timbre perception," PLoS Comput. Biol., vol.8, no.10, p.e1002759, 2012.
[30] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," IEEE Signal Process. Mag., vol.29, no.6, pp.34–43, Nov. 2012.
[31] S. Ganapathy, S. Mallidi, and H. Hermansky, "Robust feature extraction using modulation filtering of autoregressive models," IEEE/ACM Trans. Audio, Speech, Language Process., vol.22, no.8, pp.1285–1295, 2014.
[32] H. Lei, B. Meyer, and N. Mirghafori, "Spectro-temporal gabor features for speaker recognition," Proc. IEEE ICASSP, pp.4241–4244, 2012.
[33] C.C. Hsu, T.E. Lin, J.H. Chen, and T.S. Chi, "Spectro-temporal subband wiener filter for speech enhancement," Proc. IEEE ICASSP, pp.4001–4004, 2012.
[34] T.S. Chi and C.C. Hsu, "Multiband analysis and synthesis of spectro-temporal modulations of fourier spectrogram," J. Acoust. Soc. Am., vol.129, no.5, pp.EL190–EL196, 2011.
[35] 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.
[36] H. Chen and F.G. Zeng, "Frequency modulation detection in cochlear implant subjects," J. Acoust. Soc. Am., vol.116, no.4, pp.2269–2277, 2004.
[37] K. Nie, G. Stickney, and F.G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," IEEE Trans. Biomed. Eng., vol.52, no.1, pp.64–73, 2005.
[38] M. Hamouda, F. Fnaiech, and K. Al-Haddad, "A DSP based real-time simulation of dual-bridge matrix converters," Proc. IEEE ISIE, pp.594–599, 2007.
[39] S. Muller, U. Ammann, and S. Rees, "New time-discrete modulation scheme for matrix converters," IEEE Trans. Ind. Electron., vol.52, no.6, pp.1607–1615, 2005.
[40] E. Cornu, H. Sheikhzadeh, R. Brennan, H. Abutalebi, E. Tam, P. Iles, and K. Wong, "ETSI AMR-2 VAD: evaluation and ultra low-resource implementation," Proc. IEEE ICASSP, pp.II–585–8, 2003.
[41] T.F. Quatieri, Discrete-Time Speech Signal Processing Principles and Practice, Pearson Education, 2002.
[42] V. Podlozhnyuk, "FFT-based 2D convolution," NVIDIA White Paper, 2007.
[43] M. Morrone and R. Owens, "Feature detection from local energy," Pattern Recogn. Lett., vol.6, no.5, pp.303 – 313, 1987.
[44] B. Robbins and R. Owens, "2D feature detection via local energy," Image Vision Comput., vol.15, no.5, pp.353–368, 1997.
[45] M. Felsberg and G. Sommer, "The monogenic signal," IEEE Trans. Signal Process., vol.49, no.12, pp.3136–3144, 2001.
[46] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Commun., vol.9, no.4, pp.351–356, 1990.
[47] A. Varga and H.J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun., vol.12, no.3, pp.247–251, 1993.
[48] M. Aten, G. Towers, C. Whitley, P. Wheeler, J. Clare, and K. Bradley, "Reliability comparison of matrix and other converter topologies," IEEE Trans. Aerosp. Electron. Syst., vol.42, no.3, pp.867–875, 2006.
[49] ITU-T, "Objective measurement of active speech level. ITU-T Recommendation P.56," ITU-T Recommendation P.56.
[50] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of ITU-T/ETSI voice activity detectors," Proc. IEEE ICASSP, pp.1425–1428, 2001.
[51] X. Lu, M. Unoki, R. Isotani, H. Kawai, and S. Nakamura, "Adaptive regularization framework for robust voice activity detection," Proc. INTERSPEECH, pp.2653–2656, 2011.

Chung-Chien Hsu received the B.S. degree in communication engineering from National Chiao Tung University, Taiwan, in 2006. He is currently a PhD candidate in the Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan. His research interests include speech signal processing and machine learning.

Kah-Meng Cheong received the B.S. degree in communication engineering from National Chiao Tung University, Taiwan, in 2010. He is currently a PhD student in the Department of Electrical and Computer Engineering, National Chiao Tung University, Taiwan. His research interests include adaptive noise cancellation, and hardware implementation.

Tai-Shih Chi received the Ph. D. degree in electrical engineering from University of Maryland, College Park, in 2003. From August 2003 to June 2005, he was a Research Associate at the University of Maryland. He joined the Department of Electrical and Computer Engineering, National Chiao-Tung University, Taiwan, in 2005. His research interests are in neuromorphic auditory modeling, soft computing, and speech analysis.

Yu Tsao received the B.S. and M.S. degrees in Electrical Engineering from National Taiwan University in 1999 and 2001, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Georgia Institute of Technology in 2008. From 2009 to 2011, Dr. Tsao was a researcher at National Institute of Information and Communications Technology (NICT), Japan, where he engaged in research and product development in automatic speech recognition for multilingual speech-to-speech translation. Currently, he is an assistant research fellow of the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taiwan. Dr. Tsao's research interests include speech and speaker recognition, acoustic and language modeling, multimedia signal and information processing, pattern recognition and machine learning.