Neuromorphic Pitch Based Noise Reduction for Monosyllable Hearing Aid System Application

Yu-Jui Chen, Cheng-Wen Wei, Yi FanChiang, Yi-Le Meng, Yi-Cheng Huang, and Shyh-Jye Jou

***Abstract*—This paper presents a low computational complexity hardware-oriented neuromorphic pitch based noise reduction (NR) algorithm and hardware implementation for monosyllable hearing aid system applications. The proposed NR design consists of a pitch-based voice activity detector (pitch-based VAD) for speech detection and a neuromorphic noise attenuator for speech enhancement. The pitch-based VAD is developed on ANSI S1.11 based filter bank architecture and employs the characteristics of monosyllable and nonlinear energy operator (NEO) to improve the accuracy of VAD. The neuromorphic noise attenuator reduces the background noise by using the characteristics of human hearing system and the clues of speech. Simulation results show that the proposed algorithm has better SNR and PESQ performance than other non-pitch based NR algorithms in non-stationary background noise environments. Compared with multiband (mband) spectral subtraction and minimum mean square error (mmse) algorithms, the computational complexity of the proposed algorithm can save 90% computational complexity. The hardware implementation consumes 47.74 μW at 0.5 V operation with 65 nm HVT standard cell library.**

***Index Terms*—Hearing aids, Mandarin, neuromorphic, noise reduction, non-stationary, pitch.**

I. Introduction

In hearing aids (HA) systems, signals are amplified to compensate the hearing loss of patients. However, the amplified background noise may degrade the speech quality and intelligibility or even damage the residual hearing ability of patients. Thus, noise reduction is a key block in hearing aid system applications. The noise reduction algorithms based on one microphone can be categorized into three types: spectral subtraction algorithm [1], [2], statistical model based algorithm [3], [4] and subspace algorithm [5]. Although the noise reduction algorithm based on the statistical model and subspace type can efficiently suppress background noise, the computational complexity is too high to be implemented for HA applications. Thus, the spectral subtraction algorithm is frequently used in a low power HA hardware implementation. Although spectral subtraction has the superiority on computational complexity and hardware implementation, the spectral subtraction algorithm implemented in filter bank architecture might introduce artificial noise problem. This noise is mainly caused by the time-domain noise subtraction and the switches among non-linear functions of time-domain spectral subtraction algorithm.

Besides, for high de-noise efficiency and low power, noise reduction in an HA system always embeds VAD in order to distinguish between speech dominated duration and noise dominated duration. The traditional VAD usually detects voice based on energy [6], zero crossing rate [7] or entropy [8]. The computational complexity of these methods is low enough for HA applications and the accuracy is quite high in high SNR stationary noise environment, but, at the low SNR or non-stationary noise environments, the accuracy of VAD is quite low due to the inaccurate background noise estimation.

Since noise reduction algorithms use different signal processing algorithms on the noisy signal based on the indication of VAD, the performance of VAD has great impact on noise reduction performance and the power efficiency. A high performance VAD in HA systems should have the following characteristics in general: (1) High accuracy—in order to improve the speech quality and intelligibility. (2) Low computational complexity—due to limited battery power in HA systems. (3) Robust to dynamic environment—because an HA system is a portable device, and the background noise might change dramatically in the real world.

Recently, LTSV-VAD [9] uses long-term signal variability measure to discriminate noise from noisy speech and this feature is used as VAD. Due to the computation of the R frames, the computational complexity is very high and the latency incurred (300 ms) by the algorithm will exceed the latency tolerance of HA (about 10 ms $\sim$ 15 ms) [10]. Also, due to the observation of long-term signal variability, large storages are required which is not beneficial for HA systems. Hidden-Markov-model-based (HMM-based) VAD [11] shows that the accuracy is very high in spite of the low SNR, however, the computational complexity of this method is too high (Mel-frequency spectral or cepstral is required) to be applied in HA systems. For HMM-based VAD, its high accuracy is followed by high false rate. High false rate will result in preservation of noise signal in speech enhancement block. Also, due to using 20 ms frames, the latency incurred by the algorithm will exceed the latency tolerance of HA.

Finally, these two works do not show if they work in our defined non-stationary noise environment. Thus, although they have some fairly good VAD feature and may be applied in speech communication systems, they certainly are not suitable

Manuscript received January 03, 2013; revised April 15, 2013 and May 17, 2013; accepted June 04, 2013. Date of publication January 17, 2014; date of current version January 24, 2014. This research is supported in part by University shuttle program of TSMC, Industrial Technology Research Institute of Taiwan, , in part by the National Science Council under Grant NSC98-2220-E-009-008 and in part by the chip Implementation Center. This paper was recommended by Associate Editor Z. Zhang.

The authors are with the Department of Electronics Engineering and Institute of Electronics, National Chiao Tung University, Hsinchu City 30050, Taiwan (e-mail: rayraymp3@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2013.2278348

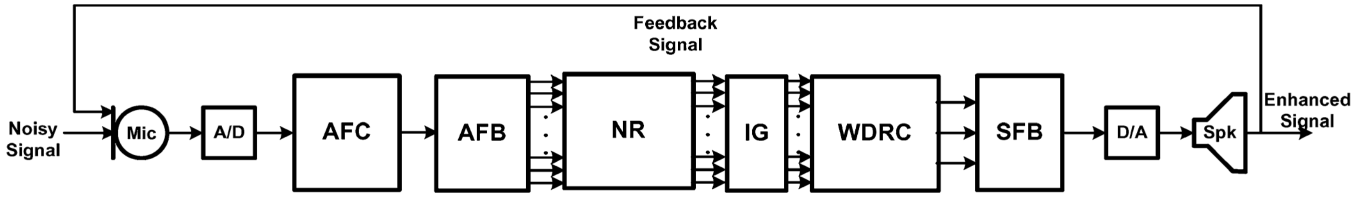

Fig. 1. Block diagram of filter bank based HA system.

to be used in HA system due to computational complexity and long latency.

Therefore, in this paper, a neuromorphic pitch-based noise reduction algorithm and design for HA systems are proposed. The proposed algorithm consists of a pitch-based voice activity detector (pitch-based VAD) and a neuromorphic noise attenuator. The proposed pitch-based VAD algorithm detects speech duration depending on the features of speech and the characteristics of monosyllable language, for example Mandarin. The proposed VAD is designed for an HA system with ANSI S1.11 [12] or low latency quasi-ANSI based filter bank architecture [13]. The filter bank used in this paper employs the subband 22nd and 39th (F22~F39) of 1/3 octave filter bank in ANSI S1.11 standard [14] which covers the most sensitive range in human hearing system. The ANSI S1.11 standard is defined to mimic the characteristic of human hearing system, whose bandwidth is narrow in low frequency and wide in high frequency subbands. In addition, a modified nonlinear energy operator is also applied in each subband to enhance the accuracy of the proposed VAD algorithm.

For speech enhancement, a neuromorphic noise attenuator is used to preserve speech and attenuate background noise depending on the characteristics of human hearing system and the onset feature of consonant. In addition, the neuromorphic noise attenuator also employs multiplication for time-domain gain smoothing to reduce the artificial noise problem of traditional spectral subtraction algorithm. Finally, considering the power limitation of HA systems and the latency tolerance of user (about 10 ms~15 ms) [10], the proposed algorithm is simplified to reduce the computational complexity. With appropriate parameter modification, the proposed algorithm is also suitable in low latency quasi-ANSI based filter bank architecture which is designed to reduce the HA system latency [13].

This paper is organized as following. Section II describes the block diagram of a filter bank based hearing aid system and the function of each block. Section III presents the details of the proposed neuromorphic noise reduction algorithm. Section IV is the simulation results in each noise environment. Section V is the hardware implementation results and Section VI is the conclusion.

II. The Process Flow of Filter Bank Based HA System

Fig. 1 shows the block diagram of a filter bank based HA system [15]. First, the voice signal collected by microphone is converted from analog to digital signal with sampling rate of 24 KHz and data word-length of 16 bits. An acoustic feedback cancellation (AFC) block reduces the feedback of the HA system. The feedback-free signal is decomposed into 18 subband signals with the ANSI S1.11 based acoustic filter bank (AFB) [12]. Then, noise reduction (NR) block attenuates the background noise and enhances the speech quality. The insertion gain (IG) block amplifies the enhanced signal of each subband individually to compensate the hearing loss of patients in each subband [16]. Wide dynamic range compressor (WDRC) block compresses the dynamic range of the amplified signal to match the residual dynamic range of patients and also protects the residual hearing ability of patients [17]. Finally, the synthesis filter bank (SFB) recombines the compressed signal of each subband.

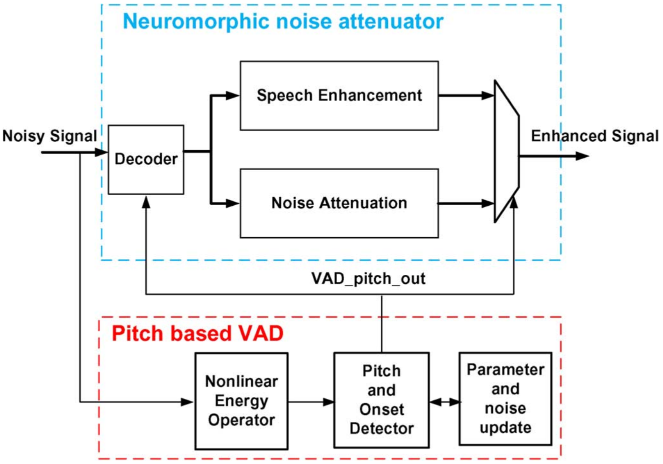

Fig. 2. The block diagram of the proposed neuromorphic pitch based noise reduction system.

III. Proposed Neuromorphic Pitch Based Noise Reduction Algorithm

Fig. 2 is the block diagram of the proposed neuromorphic pitch based noise reduction algorithm. The proposed algorithm consists of a pitch-based VAD and a neuromorphic noise attenuator. In the pitch-based VAD block, the output from filter bank is firstly processed by nonlinear energy operation to enhance the pitch and harmonic characteristics. Then pitch and onset detector are employed to help the decision of VAD. The parameters and noise amount are updated after each VAD decision. The neuromorphic noise attenuator block reduces background noise based on characteristic of human hearing perception system, masking effect and lateral inhibition effect in cochlear. Depending on the VAD result, speech enhancement or noise attenuation block processes the noisy speech or noise signal respectively to produce the enhanced signal.

Firstly, this section shows the characteristics of monosyllable language and human hearing system. Secondly, the design principle of the pitch-based VAD and neuromorphic noise attenuator are addressed. Finally, for HA hardware implementation, the proposed algorithm is modified to reduce the computational complexity.

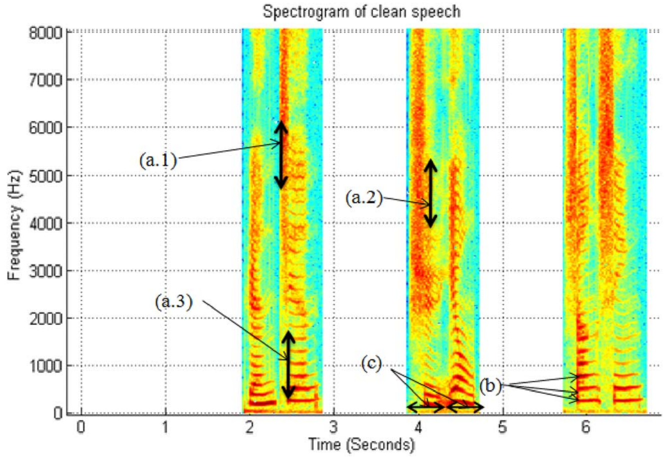

Fig. 3. Spectrogram of clean speech. (a.1) onset of consonant. (a.2) offset of consonant. (a.3) onset of vowel. (b) pitch and corresponding harmonics. (c) length of each word in monosyllable language.

A. Characteristics of Monosyllable Language and Human Hearing System

Fig. 3 shows the spectrogram of speech signal. The indications in the figure illustrate several important features of monosyllable language (in this case, Mandarin), which is described as following [18], [19]:

1) Onset and Offset Features: (a.1) and (a.3) show that consonant or vowel in each subband occurs almost at the same time. Also, the termination time of consonant in each subband is almost the same, as indication (a.2) shows.

2) Pitch and Corresponding Harmonics: Indication (b) shows that the spectrogram of vowel is composed of a fundamental frequency F0, called pitch, and the corresponding harmonics of F0. The pitch and harmonics feature is due to the vibration of the vocal cords during speaking and is the most important feature of speech detection in low frequency. Of course, the pitch and corresponding harmonics also have the onset and offset features described in (a).

3) Length of Each Word in Monosyllable Language: Language like English has monosyllables and polysyllables, which means the length of time of each word is different. Unlike English, the length of time of each monosyllable language is almost the same for the same speaker, as indication (c) shows.

The characteristics of human hearing system are described as following:

4) Masking Effect: The human hearing system is like a frequency discriminator. However, if two or more sound sources are close to each other, either in frequency or time domain, the one with the highest energy will raise the threshold of audibility of other low energy sound sources. The highest energy sound source is called the masker and those inaudible low energy sound sources are called the masked. Whenever the magnitude of masked sound is smaller than the masking threshold caused by the masker sound, the masked sound is not audible to human hearing system. Moreover, the closer to the masker component the masked component is, the higher the masking threshold is. This phenomenon is called masking effect [20].

Because the speech concentrates on particular subbands, under the situation that the segmental SNR ($\mathrm{SNR_{seg}}$) is high enough, it is reasonable to assume that the background noise is masked by speech and do not need to be processed in those pitch and corresponding harmonics dominated subbands.

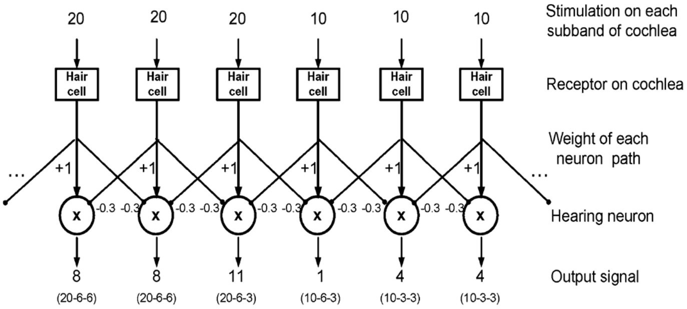

Fig. 4. Lateral inhibition effect in human hearing system.

5) Lateral Inhibition Effect: In the cochlea of human hearing system, the phenomenon of lateral inhibition effect is that a subband with high energy component will enhance itself and inhibits the neighboring low energy subbands. Therefore, the activity of each subband in cochlea is decided not only by the energy of the target subband but also by the neighboring ones. As shown in Fig. 4, although there are only two kinds of input amplitudes, the output responses of neurons are different because of the lateral inhibition effect, especially at the edge of the two magnitudes.

The lateral inhibition effect was used for speech enhancement and recognition algorithms [21]. Because the speech concentrates on particular subbands, the speech tones can inhibit background noise of neighbor subbands under the assumption that the segmental SNR is high enough.

B. Pitch Based Voice Activity Detector

The proposed pitch-based VAD algorithm is developed based on ANSI S1.11 standard 1/3 octave filter bank architecture [12]. The unique feature is to use pitch and onset feature of speech to detect speech duration. In addition, nonlinear energy operator (NEO), phoneme keeper and majority vote are also adopted to improve the VAD accuracy. Since the pitch and its harmonics mainly locate at low frequency subbands, we consider subbands F22 to F30 for pitch and onset detection.

1) Nonlinear Energy Operator: The nonlinear energy operation is a general method in neuron signal processing to highlight the peak of neuron spike to further separate it from noise and DC level in the time domain [22].Therefore, the proposed pitch-based VAD adopts a modified nonlinear energy operator in each subband, as shown in (1).

$$\begin{cases} SNR_{NEO}(i,j) \\ \quad = \frac{P_{avg}(i,j)^2}{P_{avg}(i,j+1)^2} & \text{for } j = 22 \\ SNR_{NEO}(i,j) \\ \quad = \frac{P_{avg}(i,j)^2}{P_{avg}(i,j-1) \times P_{avg}(i,j+1)} & \text{for } j = 23 \sim 30 \end{cases} \tag{1}$$

where

$$P_{avg}(i,j) = \frac{1}{K_1}\frac{1}{S(j)}\sum_{h=i-(K_1-1)}^{i}\sum_{k=1}^{S(j)} Data_{in}(h,j,k)^2$$

where $Data_{in}(i,j,k)$ is the magnitude of sample k, i is the frame index, j is the target subband and $S(j)$ is the number of sample at the target subband. $P_{avg}(i,j)$ is the average power of $i-(K1-1)$ to i frames. K_1 is set as 16 corresponding to 21.3 ms which can be considered stationary for analysis [23]. $SNR_{NEO}(i,j)$ is the result of NEO process of subband j in frame i. Eqn (1) processes the decomposed signal of low frequency subband F22 to F30 simultaneously to emphasize those subbands dominated by pitch or harmonics.

2) Pitch Detector and Onset Detector of Speech: Whenever pitch and its harmonics occur, the average power of corresponding subbands will increase quickly. For example, assume pitch occurs at subband F23 with center frequency at 260 Hz. Because the filter bank is 1/3 octave based, the second harmonic will locate at subband F26 and the fourth harmonic will locate at subband F29. Therefore, the average power of subbands F23, F26, and F29 will become larger than other noise dominated subbands. Although in some situations, pitch may locate at the overlap range of two neighbor subbands or the third harmonic locates at the neighbor subband of the second or fourth harmonic, causing the power of these neighbor subbands to be quite the same. Fortunately, the NEO is a bi-sides operation, which means it will consider the average power of subbands at both sides of the target subband. Therefore, with appropriate threshold value, the NEO operator can still separate the pitch dominated subbands from noise dominated subbands in these situations. Therefore, $SNR_{NEO}(i,j)$ of pitch dominated subbands become very large due to the modified nonlinear energy operation. Thus, the proposed VAD employs pitch detector (2) and onset detector (3) to decide the temporal result of pitch-based VAD, $VAD_{pitch\_temp}(i)$. First, the pitch detector, $Detect_{pitch}(i)$, will be set to 1 if at least one of the pitch detection criteria is satisfied, meaning the pitch and corresponding harmonics are detected.

$$Detect_{pitch}(i)=\begin{cases}1, & Pitch1\cup Pitch2\\ & \quad\cup Pitch3==1 \qquad \text{for } j=22,23,24\\ 0, & \text{Otherwise}\end{cases} \tag{2}$$

where

$$\begin{aligned}Pitch1 &= (SNR_{NEO}(i,j) > Thr_{p1}(i))\\ &\quad \cap (SNR_{NEO}(i,j+3) > Thr_{p2}(i))\\ Pitch2 &= (SNR_{NEO}(i,j+3) > Thr_{p2}(i))\\ &\quad \cap (SNR_{NEO}(i,j+6) > Thr_{p3}(i))\\ Pitch3 &= (SNR_{NEO}(i,j) > Thr_{p1}(i))\\ &\quad \cap (SNR_{NEO}(i,j+6) > Thr_{p3}(i))\end{aligned}$$

where $Thr_{p(j)}(i)$ is the threshold of pitch detector of the ith frame for subband j. The operator '$\cup$' implies logical OR and '$\cap$' implies logical AND. The relational operator '==' checks if the values of two operands are equal or not and '>' checks if the value of left operand is greater than the value of right operand. Due to different bandwidth among subbands in ANSI S1.11 standard, $SNR_{NEO}(i,j)$ is larger for lower subbands as compared to higher subbands. So in pitch detection, we partition subbands 22 to 30 into three segments. Each segment has its own threshold which is higher for lower segment and lower for higher segment. Then, another speech characteristic, onset feature, is embedded into VAD equation. The onset detector, $Detect_{onset}(i)$, means that only the appearance of harmonic pattern is not enough. These harmonics must appear simultaneously to match the onset characteristic of speech, as shown in (3). $M_{frame}(i,j)$ is the average magnitude of subband j in frame i. $M_{avg\_low}(i,j)$ is the average magnitude in the previous K_2 frames of subband j and $M_{longterm}(i,j)$ is the long term average magnitude of subband j. It is the convex combination between $M_{longterm}(i-1,j)$ and $M_{frame}(i,j)$ for tracking the long term variation of magnitude for subband j. $Thr_{on(j)}$ is the threshold of onset detector of the target subband j. The choice of $Thr_{on(j)}$ is similar to pitch detector. K_2 is the same with K_1 and the larger the K_3 is, the smoother the long term average magnitude is. K_3 is set as 6 empirically.

$$Detect_{onset}(i)=\begin{cases}1, & Onset1\cup Onset2\\ & \quad\cup Onset3==1 \qquad \text{for } j=22,23,24\\ 0, & \text{Otherwise}\end{cases} \tag{3}$$

where

$$\begin{aligned}Onset1 &= \left(\frac{M_{avg\_low}(i,j)}{M_{longterm}(i,j)} > Thr_{on1}\right)\\ &\quad \cap \left(\frac{M_{avg\_low}(i,j+3)}{M_{longterm}(i,j+3)} > Thr_{on2}\right)\\ Onset2 &= \left(\frac{M_{avg\_low}(i,j+3)}{M_{longterm}(i,j+3)} > Thr_{on2}\right)\\ &\quad \cap \left(\frac{M_{avg\_low}(i,j+6)}{M_{longterm}(i,j+6)} > Thr_{on3}\right)\\ Onset3 &= \left(\frac{M_{avg\_low}(i,j)}{M_{longterm}(i,j)} > Thr_{on1}\right)\\ &\quad \cap \left(\frac{M_{avg\_low}(i,j+6)}{M_{longterm}(i,j+6)} > Thr_{on3}\right)\end{aligned}$$

with

$$\begin{aligned}M_{avg\_low}(i,j) &= \frac{1}{K_2}\sum_{h=i-(K_2-1)}^{i} M_{frame}(h,j)\\ M_{longterm}(i,j) &= M_{longterm}(i-1,j)\times(1-2^{-K_3})\\ &\quad + M_{frame}(i,j)\times 2^{-K_3}\\ M_{frame}(i,j) &= \frac{1}{S(j)}\sum_{k=1}^{S(j)} Data_{in}(i,j,k)\end{aligned}$$

Whenever one of the three pitch detector conditions is satisfied (case $j\ =\ 22$ or 23 or 24), the pitch-based VAD will check whether the corresponding onset condition is also satisfied or not. If both conditions are confirmed, the temporal result of VAD, $VAD_{pitch\_temp}(i)$, will be set to 1. Because the shape of the oral tract will affect the distribution of the average power among subbands, the power of subband F26 may be smaller than power of subband F29. By adding the pitch and onset detection criteria, $Pitch3$ and $Onset3$, the case can be covered. Other special case such as power only located in one subband can be solved by examining if there is large power concentrated in one subband when (2) and (3) are not satisfied.

It is worth to mention that unlike the traditional noise estimation, the function of $M_{longterm}(i,j)$ is independent of the result of VAD. Therefore, the noise estimation is not affected by the performance of VAD and environment variation.

3) Phoneme Keeper and Majority Vote: Since the length of time of each monosyllable word is almost the same for a speaker, a phoneme keeper $(Detect_{phoneme})$ uses this feature to keep the continuity of monosyllable word and enhance the accuracy of VAD. In addition, a majority vote scheme is also employed to smooth the result of VAD and avoids the disturbance caused by short time peak noise. Fig. 5 shows the flow chart of the proposed pitch-based VAD and (4) shows the equation for $Detect_{phoneme}(i)$ and $Majority(i)$.

$$Detect_{phoneme}(i)=\left(\bigcap_{k=1}^{K_6} VAD_{pitch}(i-k)\right)\cap(keeper\geq 1) \quad (4a)$$

$$Majority(i)=\sum_{j=i-(K_4-1)}^{i} VAD_{pitch\_temp}(j) \quad (4b)$$

The operator '$\geq$' checks if the value of left operand is greater than or equal to the value of right operand. $Detect_{phoneme}(i)$ is valid whenever the speech is detected and stable for a span of frames (K_6) which is close to the length of time of a phoneme (about 10 ms). In this case, K_6 is set to be 6. The output of VAD, $VAD_{pitch}(i)$, will be kept at 1 for the time of single monosyllable character which is about 300 ms. The $keeper$ and $length_{keeper}$ are used to control the duration of phoneme keeper and $Majority(i)$ is the result of majority vote. The $Majority(i)$ is decided by calculating the number of 1 of temporal result of VAD, $VAD_{pitch\_temp}(i)$, in the previous K_4 frames. If $Majority(i)$ is larger than K_5, the $VAD_{pitch}(i)$ will be set to 1. Otherwise, $VAD_{pitch}(i)$ will be set to 0. K_4 and K_5 (roughly equals to 0.5 times K_4) are set as 12 and 6 respectively in this experiment and K_5 is a trade-off factor. With the phoneme keeper and majority vote, the proposed VAD can keep the speech continuality and filter out the short time peak noise disturbance.

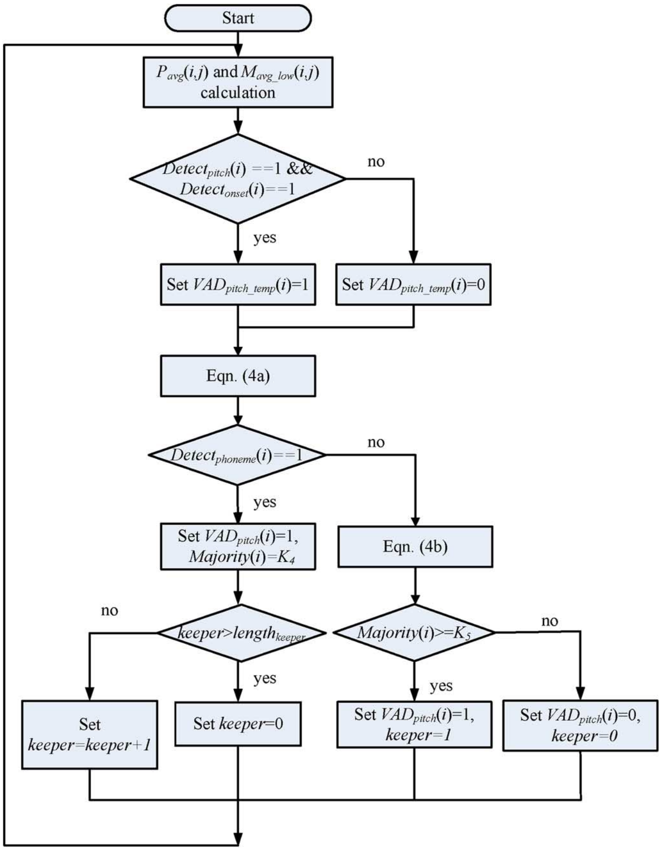
Start
$P_{avg}(i,j)$ and $M_{avg\_low}(i,j)$ calculation
$Detect_{pitch}(i)$ ==1 && $Detect_{onset}(i)$==1
no
yes
Set $VAD_{pitch\_temp}(i)$=1
Set $VAD_{pitch\_temp}(i)$=0
Eqn. (4a)
$Detect_{phoneme}(i)$==1
no
yes
Set $VAD_{pitch}(i)$=1, $Majority(i)$=K_4
Eqn. (4b)
no
$keeper$>$length_{keeper}$
yes
$Majority(i)$>=K_5
no
yes
Set $keeper$=$keeper$+1
Set $keeper$=0
Set $VAD_{pitch}(i)$=1, $keeper$=1
Set $VAD_{pitch}(i)$=0, $keeper$=0

Fig. 5 Flow chart of the proposed pitch based VAD algorithm.

C. Neuromorphic Noise Attenuator

The proposed neuromorphic noise attenuator (NNA) algorithm shown in Fig. 6 enhances speech intelligibility based on the characteristic of human hearing system. That is, the noise in the speech dominated subband is not necessary to be processed when the $\mathrm{SNR_{seg}}$ is above or equals to 0 owing to the masking effect principle. Moreover, the subbands processed by attenuator are partitioned into groups of low frequency subbands and high frequency subbands. Finally, in order to protect the short-term voice and the continuity of speech, a short voice protector (SVP) mechanism is also applied into the neuromorphic noise attenuator algorithm.

First of all, if $VAD_{pitch}(i)$ is 0 and the current frame is not detected as short voice protection zone, which means $Majority(i) < K_7$, the current frame is processed by noise attenuation block and the gain is set to $Gain_{atte}$, 0.25 in our case, for background noise attenuation. The neuromorphic noise attenuator reduces background noise by using multiplication instead of subtraction to prevent the artificial noise problem of time-domain spectral subtraction.

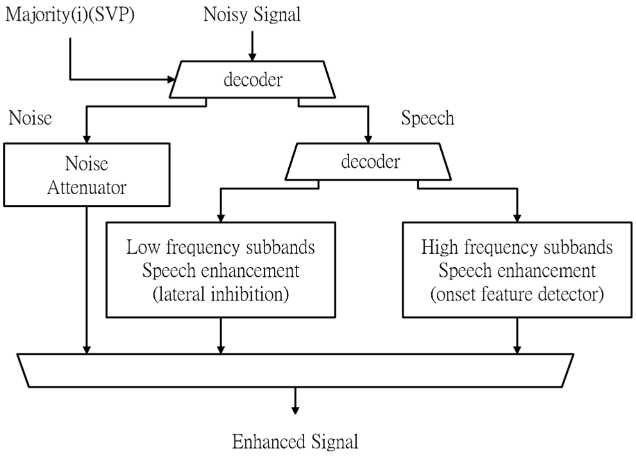
Majority(i)(SVP)
Noisy Signal
decoder
Noise
Speech
Noise Attenuator
decoder
Low frequency subbands Speech enhancement (lateral inhibition)
High frequency subbands Speech enhancement (onset feature detector)
Enhanced Signal

Fig. 6. The block diagram of the proposed neuromorphic noise attenuator algorithm.

On the other hand, when the current frame is detected as speech, the low frequency subbands and high frequency subbands use different approaches to do speech enhancement. For low frequency subbands, the energy of speech is concentrated on several subbands, especially the subbands with pitch and corresponding harmonic tones. Therefore, low frequency subbands, F22 to F30, shall do pitch and corresponding harmonic tones analysis. Thus, the proposed neuromorphic noise attenuator algorithm employs the lateral inhibition principle to attenuate the background noise and preserves the speech signal in low frequency subbands, as shown in (5).

$$\begin{cases} NNA_{lowf}(i,j)=\frac{P_{avg}(i,j)^2}{P_{avg}(i,j+1)^2} & \text{for } j=22 \\ NNA_{lowf}(i,j)=\frac{P_{avg}(i,j)^2}{P_{avg}(i,j-1)\times P_{avg}(i,j+1)} & \text{for } j=23\sim 30 \end{cases} \tag{5}$$

The lateral inhibition effect makes the subband with high energy be boosted and attenuates the energy of neighbor subbands, and vice versa. In addition, the energy of target subband is squared in order to normalize the $NNA_{lowf}(i,j)$. It is worth to mention that the equation of $NNA_{lowf}(i,j)$ is identical to (1) for modified nonlinear energy operation. Thus, the computational complexity can be reduced by merging these two operations.

Then, the gain of subband F22 to F30, $Gain\_L(i,j)$, is decided by (6).

$$Gain\_L(i,j)=\frac{1}{K_8}\sum_{h=i-(K_8-1)}^{i} Gain\_L_{temp}(h,j) \qquad \text{for } j=22\sim 30 \tag{6}$$

where

$$\begin{aligned} &Gain\_L_{temp}(i,j) \\ &\quad=\begin{cases} \mathrm{Gain}_{enh\_lowf}, & (Majority(i)\geq K_7) \\ & \quad\cap(NNA_{lowf}(i,j)\geq Thr_{lowf}(j)) \\ \mathrm{Gain}_{atte2}, & (Majority(i)\geq K_7) \\ & \quad\cap(Thr_{lowf}(j)>NNA_{lowf}(i,j)) \end{cases} \\ &Thr_{lowf}(j) \\ &\quad=G_{preserve}\times Thr_{p(j)}(i) \quad \text{for } j=22\sim 30 \end{aligned}$$

The temporal gain of target subband, $Gain\_L_{temp}(i,j)$, will be set to $Gain_{enh\_lowf}$, whenever the $NNA_{lowf}(i,j)$ is higher than threshold value, $Thr_{lowf}(j)$. This situation means that the current subband is speech and the target subband is dominated by pitch or corresponding harmonic tones. Hence the signal of target subband should be preserved. Otherwise, the temporal gain of target subband will be set to $Gain_{atte2}$. Because this situation means although the current frame is speech, the target subband is noise dominated and should be attenuated. K_7 and K_8 are constants. The neuromorphic noise attenuator algorithm averages the $Gain\_L_{temp}(i,j)$ of current frame with the previous K_8-1 frames to produce the final gain, $Gain\_L(i,j)$, of the target subband. K_7 is set as 3, one fourth of K_4, and K_8 is chosen as 4 to smooth gain in our case. $Gain_{enh\_lowf}$ and $Gain_{atte2}$ are 1 and 0.25 respectively to simplify hardware implementation. In addition, to preserve the speech and reduce distortion during speech duration, $G_{preserve}$ is designed to be smaller than 1 so that $Thr_{lowf}(j)$ is smaller than $Thr_{p(j)}(i)$.

In high frequency subbands, F31 to F39, the speech components in these subbands are consonants instead of vowels. Therefore, the neuromorphic noise attenuator algorithm uses the most obvious characteristic of consonant, the onset feature, as the gain index in high frequency subbands, as shown in (7).

$$NNA_{highf}(i,j)=\frac{M_{avg\_high}(i,j)}{M_{longterm}(i,j)} \text{ for } j=31\sim 39 \tag{7}$$

where

$$\begin{aligned} M_{avg\_high}(i,j)&=\frac{1}{K_9}\sum_{h=i-(K_9-1)}^{i} M_{frame}(h,j) \\ M_{longterm}(i,j)&=M_{longterm}(i-1,j)\times(1-2^{-K_{10}}) \\ &\quad+M_{frame}(i,j)\times 2^{-K_{10}} \\ M_{frame}(i,j)&=\frac{1}{S(j)}\sum_{k=1}^{S(j)} Data_{in}(i,j,k) \end{aligned}$$

where $M_{frame}(i,j)$ is the average magnitude of subband j in frame i. $M_{avg\_high}(i,j)$ is the average magnitude from $i-k_9+1$ to i frames of subband j. K_9 and K_{10} are constants and the principle to select the values of these two constants is the same with K_2 and K_3. Eqn. (7) means that the $NNA_{high}(i,j)$ of the target high frequency subband is proportional to the slope of magnitude. For the target subband with high slope of magnitude, it might indicate the onset feature of a new component has been detected and the neuromorphic noise attenuator will provide corresponding gain to the target subband depending on its $NNA_{high}(i,j)$, as shown in (8).

$$Gain\_H(i,j)=\frac{1}{K_8}\sum_{h=i-(K_8-1)}^{i} Gain\_H_{temp}(h,j) \qquad \text{for } j=31\sim 39 \tag{8}$$

where (See equation at bottom of page)

The neuromorphic noise attenuator algorithm averages the $Gain\_H_{temp}(i,j)$ of the current frame with previous K_8-1 frames to produce the final gain, $Gain\_H(i,j)$, of the target subband. The starting duration of speech and the short duration speech might be missed because it always takes time to react the pitch or onset feature on the average magnitude or power. Therefore, this part of speech will be attenuated before the $VAD_{pitch}(i)$ switches to 1, which may influence the speech quality and intelligibility. In order to improve this situation, the neuromorphic noise attenuator has a short voice protection zone. Whenever $K_5>Majority(i)\geqq K_7$, although

$$Gain\_H_{temp}(i,j)=\begin{cases} Gain_{enh\_highf1}, & (Majority(i)\geq K_7)\cap(NNA_{highf}(i,j)\geq Thr_{high1}) \\ Gain_{enh\_highf2}, & (Majority(i)\geq K_7)\cap(Thr_{high1}>NNA_{highf}(i,j)\geq Thr_{high2}) \\ Gain_{enh\_highf3}, & (Majority(i)\geq K_7)\cap(Thr_{high2}>NNA_{highf}(i,j)\geq Thr_{high3}) \\ Gain_{atte2}, & (Majority(i)\geq K_7)\cap(Thr_{high3}>NNA_{highf}(i,j)) \end{cases}$$

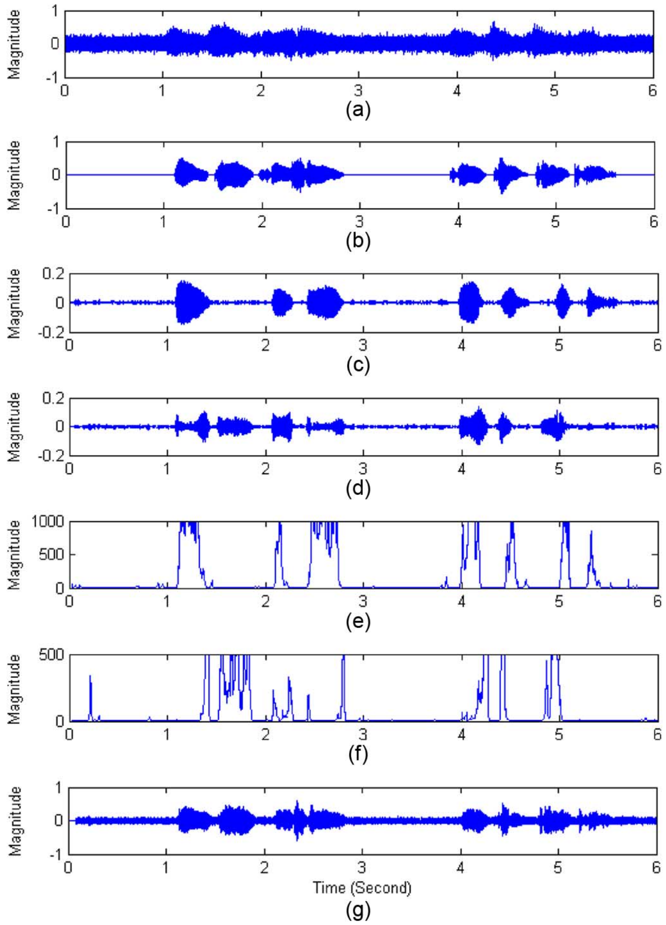

Fig. 7. Simulation results of noisy speech in $\mathrm{SNR_{seg}}$ 0 dB white noise environment. (a)Waveform of noisy speech. (b) Waveform of clean speech. (c) The decomposed input signal of subband F23. (d) The decomposed input signal of subband F24. (e) The result of $\mathrm{SNR_{NEO}}(\mathrm{i}, 23)$. (f) The result of $\mathrm{SNR_{NEO}}(\mathrm{i}, 24)$. (g) The enhanced speech.

the $VAD_{pitch}(i)$ is 0, the signal will be detected as short voice protection zone and processed by speech enhancement block instead of noise attenuation block. $Gain_{enh\_highf1}$, $Gain_{enh\_highf2}$, and $Gain_{enh\_highf3}$ are set as 1, 0.75, 0.5, respectively for hardware implementation. After the gain is decided, the threshold can be decided empirically.

The latency incurred by the proposed algorithm is only 1 frame which is 32 samples and corresponds to only 1.3 ms. The data utilized by the average operation in (1) to (8) are previous information. Thus, there is no latency introduced by this operation.

Fig. 7 shows the simulation results of the noisy speech in $\mathrm{SNR_{seg}}$ 0 dB white noise environment. Fig. 7(c), (d) are the original decomposed input signals of subband F23 and F24, and Fig. 7(e), (f) are the results of SNR_{NEO} of subband F23 and F24. As shown in the figure, the fundamental frequency of the pitch are in subband F23 and F24 during speech. Through the NEO process, the difference between speech dominated duration and noise dominated duration can be greatly enhanced to improve the accuracy of VAD. Fig 7(g) is the waveform of enhanced speech after the neuromorphic noise attenuator algorithm.

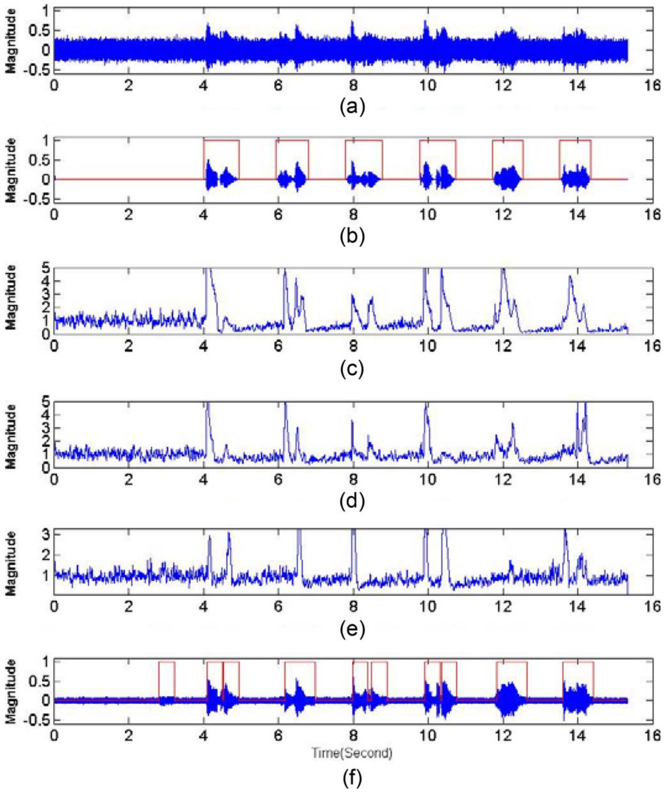

Fig. 8. Simulation results of the proposed algorithm and onset detector in $\mathrm{SNR_{seg}}$ 0 dB white noise environment. (a)Waveform of noisy speech (b) Waveform of clean speech and ideal VAD signal. (c) The result of onset detector of subband F23, $\mathrm{M_{avg\_low}(i, 23)/M_{longterm}(i, 23)}$. (d) The result of onset detector of subband F26, $\mathrm{M_{avg\_low}(i, 26)/M_{longterm}(i, 26)}$. (e) The result of onset detector of subband F29, $\mathrm{M_{avg\_low}(i, 29)/M_{longterm}(i, 29)}$. (f) The enhanced speech after the neuromorphic noise attenuator algorithm and the results of the pitch based VAD algorithm.

Fig. 8 shows results of the proposed VAD algorithm and onset detector in $\mathrm{SNR_{seg}}$ 0 dB white noise environment. Fig. 8(b) is the clean speech and the ideal VAD signal. Fig. 8(c)–(e) are the results of onset detector in three subbands, F23, F26 and F29. As shown in the figure, the slope of magnitude increases rapidly during the start-up of the speech. Fig. 8(f) is the enhanced speech by using the neuromorphic noise attenuator algorithm and the results of the proposed pitch-based VAD algorithm, as we can see, the background noise is attenuated and the speech is preserved due to precise results of VAD.

D. Low Power Architecture Design

In the proposed algorithm, multiplications are the bottleneck of the computational complexity. There are three major usage of multiplication: the NEO operation of low frequency subbands, the average power calculation of subband F22 to F31 and the output signal multiplication of each subband. Without degrading the performance seriously, the proposed VAD and neuromorphic noise attenuator algorithms are modified to reduce the computational complexity.

First of all, the average power of subband F22 to F31 in (1) are modified from mean of square to average of magnitude in order to reduce the number of multiplication and the bit-width of the multiplier. In addition, (3) for $Detect_{onset}(i)$ needs division operations in each low frequency subband. However, a division operation has a higher hardware complexity as compared with a multiplication operation. In order to avoid division operations,

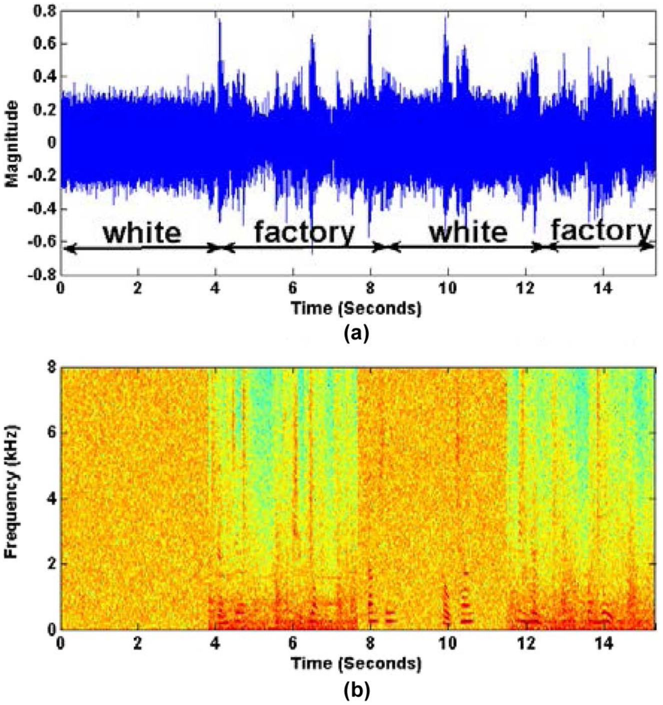

Fig. 9. (a) Waveform of white (0 dB)—factory (0 dB) test pattern (b) Spectrogram of white (0 dB)—factory (0 dB) test pattern.

the denominator is transposed to the other side of the equation. After this modification, these equations can avoid the usage of division operation which can save many hardware resources and does not affect the performance.

Finally, each frame in the proposed algorithm has 32 samples, so the number of sample after ANSI S1.11 analysis filter bank becomes 189 $(1+1+1+2+2+2+4+4+4+8+8+8+16+16+16+32+32+32)$, which means there are 189 multiplication operations just for output signal calculation in a frame.

The computational complexity of this part is even higher than spectral subtraction algorithm. Therefore, the proposed algorithm employs a shift-and-add operator with 0.125 resolution to replace the multiplication. With this tradeoff between flexibility and complexity, 1 multiplication operation can be replaced by about 1.5 shift operation and 1 addition operation in average.

IV. Simulation Results and Comparisons

A. Simulation Results

Simulations are executed to evaluate the performance of the proposed pitch-based VAD and neuromorphic noise attenuator algorithm. For Mandarin speech database, 27 Mandarin 2-characters and 10 Mandarin sentences are utilized. Among the speech frames, 33 % and 67% of the frames is high frequency sounds (more than 50% of the sound energy is located above 1000 Hz) and low frequency sounds (more than 50% of the sound energy is located below 1000 Hz), respectively. Both stationary and non-stationary noise environments are applied to the simulations. There are four types of background noises in the stationary noise environment, white, factory, car and babble provided by NOISEX-92 database [24]. Each type of noise is added to speech to have 11 different $\mathrm{SNR_{seg}}$ levels (0 dB $\sim$ 10 dB). On the other hand, the non-stationary noise environment is with different kinds of the background noises (like the situation that people moves from indoor to outdoor in real world). There are three kinds of non-stationary background noises, white-factory, white-car and white-babble. These non-stationary noises are also added to speech to have five various $\mathrm{SNR_{seg}}$ levels (0-0 dB, 0-3 dB, 0-5 dB, 0-7 dB, 0-10 dB). The white (0 dB)—factory (0 dB) situation means the background noise changes dramatically between $\mathrm{SNR_{seg}}$ 0 dB white noise and $\mathrm{SNR_{seg}}$ 0 dB factory noise during speech, as shown in Fig. 9.

In the following performance simulations, the accuracy of the proposed pitch-based VAD (VAD_accuracy) is compared with the ideal VAD signal (manually labeled). VAD_accuracy is the percentage of the number of correctly detected frames to total number of frames. The number of correctly detected frames is total number of frames minus the number of missing frames (speech frame but detected as noise frame) and false alarmed frames (noise frame but detected as speech frame). For the proposed neuromorphic noise attenuator algorithm, SNR, segmental SNR $(\mathrm{SNR_{seg}})$ and perceptual evaluation of speech quality (PESQ) [25] indices are used for the performance evaluation. PESQ is the index of objective measurement based on the speech quality. Eqn. (9) shows the definition of SNR and $\mathrm{SNR_{seg}}$,

$$\mathrm{SNR} = \frac{Energy_{cleanspeech}}{Energy_{noise}} \tag{9a}$$

$$\mathrm{SNR_{seg}} = \frac{Energy_{cleanspeech}}{Energy_{noise}} \text{ for } VAD_{idea} = 1 \tag{9b}$$

where $Energy_{cleanspeech}$ is the energy of clean speech and the $Energy_{noise}$ is the energy of the noise. $\mathrm{SNR_{seg}}$ reflects the performance of NR during speech duration and the SNR reflects the performance of NR in overall duration.

The accuracy of the proposed pitch-based VAD algorithm is compared with SNR based [6] and entropy based [8] VAD algorithm. Note that SNR based VAD algorithm utilizes 512-point FFT of which the frequency resolution is higher than 18-subband ANSI filter bank employed by entropy based VAD algorithm. The speech enhancement performance of the proposed algorithm is also compared with diverse types of algorithms such as multiband (mband) spectral subtraction [2], wiener filter (wiener_as, wiener_wt) [26], [27], minimum mean square error (mmse) [3] and subspace (klt) [5] algorithms. The programs of these compared algorithms are provided by [23].

In average, the VAD accuracy of the modified algorithm for low power architecture design is about 3% lower than the original algorithm in stationary noise environment and about 4% lower in non-stationary noise situation. With the slight degradation of the performance, we can reduce the computational complexity dramatically for low power consideration as will be shown in Table IV. The following results are presented by the original algorithm.

Table I and Table II show the accuracy of the proposed pitch-based VAD algorithm in stationary noise environment and non-stationary noise environment. The accuracy of the proposed VAD is above 85 % at the white, factory and car noise. This is because the pitch feature is basically independent of these noise environments. However, the accuracy in babble

TABLE I
THE ACCURACY OF THE PROPOSED PITCH BASED VAD ALGORITHM IN STATIONARY NOISE ENVIRONMENTS

For word cases				
SNR_{seg} (dB)	VAD_Accuracy (%)			
	White	Factory	Car	Babble
0	88.86	83.95	86.88	66.45
3	88.83	86.58	86.89	69.87
5	89.94	89.48	86.90	69.99
7	89.94	88.44	86.82	72.62
10	90.05	89.97	85.87	76.43
For sentence cases				
SNR_{seg} (dB)	VAD_Accuracy (%)			
	White	Factory	Car	Babble
0	86.99	83.91	87.94	66.12
3	88.19	83.95	89.06	69.79
5	87.76	84.87	88.47	72.10
7	87.84	85.79	88.35	73.67
10	88.23	86.44	88.68	75.70

TABLE II
THE ACCURACY OF THE PROPOSED PITCH BASED VAD ALGORITHM IN NON-STATIONARY NOISE ENVIRONMENTS

For word cases			
SNR_{seg} (dB)	VAD_Accuracy (%)		
	White-Factory	White- Car	White-Babble
0-0	85.36	88.11	76.56
0-3	86.48	88.12	78.21
0-5	87.46	88.65	81.11
0-7	88.69	88.61	82.11
0-10	88.81	88.31	82.68
For sentence cases			
SNR_{seg} (dB)	VAD_Accuracy (%)		
	White-Factory	White- Car	White-Babble
0-0	80.23	87.63	73.20
0-3	82.78	88.03	73.54
0-5	84.24	88.09	75.75
0-7	86.03	87.49	77.51
0-10	87.01	87.58	80.19

noise is around 70 % since there are several pitch components in the background noise. Therefore, the number of false alarm frames in babble noise environment is much higher than that of other background noise environments. Another characteristic of the proposed VAD is the accuracy is basically independent of the $\mathrm{SNR_{seg}}$, which is also due to that the pitch is independent of background noise. Therefore, the proposed VAD algorithm can have high accuracy in low $\mathrm{SNR_{seg}}$ noise environment. This is very important since a hearing loss person needs better SNR as compared to a normal hearing person. According to [28], the average SNR deficit of mild hearing loss people is about 4 dB. For a normal hearing person, a speech with SNR of 5 dB can be well recognized. In addition, the accuracy of the proposed VAD is almost no degradation in non-stationary noise environment, which means the proposed VAD algorithm works well at the dramatic change of background noise.

Table III shows accuracy of the proposed pitch-based VAD algorithm with the phoneme keeper procedure for all the background noise cases. The accuracy of the proposed VAD can be improved about 20% with the phoneme keeper procedure which means the phoneme keeper improves the VAD accuracy effectively.

TABLE III
THE ACCURACY IMPROVEMENT OF THE PROPOSED PITCH BASED VAD ALGORITHM BY PHONEME KEEPER IN STATIONARY AND NON-STATIONARY NOISE ENVIRONMENTS

SNR_{seg} (dB)	VAD Accuracy (%)			
	Stationary		Non-Stationary	
	w/o PK	w/i PK	w/o PK	w/i PK
0	63.68	81.54	65.05	83.34
3	64.84	83.04	65.91	84.27
5	65.55	84.08	66.46	85.74
7	65.87	84.46	66.95	86.47
10	66.43	85.58	67.60	86.60

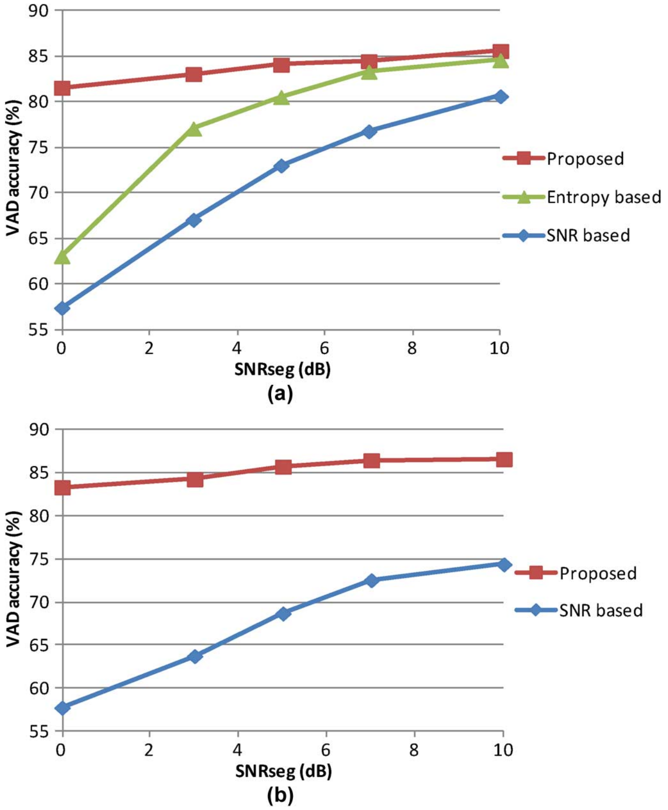

Fig. 10. Comparison of the VAD accuracy in (a) stationary noise environments and (b) non-stationary noise environments.

Fig. 10 shows the VAD accuracy of the proposed algorithm and other algorithms [6], [8] in stationary and non-stationary noise environments. In low SNR case, the accuracy of the pitch based algorithm is about 20% higher in stationary and 25% higher in non-stationary noise environment than the other algorithms. Note that in non-stationary noise environment, there is no simulation result of entropy based VAD algorithm. So, only two curves are shown in Fig. 10(b).

For the performance of neuromorphic noise attenuator, Fig. 11 and Fig. 12 show the average improvement in SNR, $\mathrm{SNR_{seg}}$ and enhanced PESQ of the proposed algorithm and the other algorithms. In the four types of stationary noise environments, the proposed algorithm performs better than mband, wiener_wt algorithm and comparable to klt subspace algorithm in low $\mathrm{SNR_{seg}}$ noise environment. More important, at 0 dB, the SNR/$\mathrm{SNR_{seg}}$ improvement is 8.0/5.5 dB which can compensate the deficit of mild hearing loss patients. Although the $\mathrm{SNR_{seg}}$

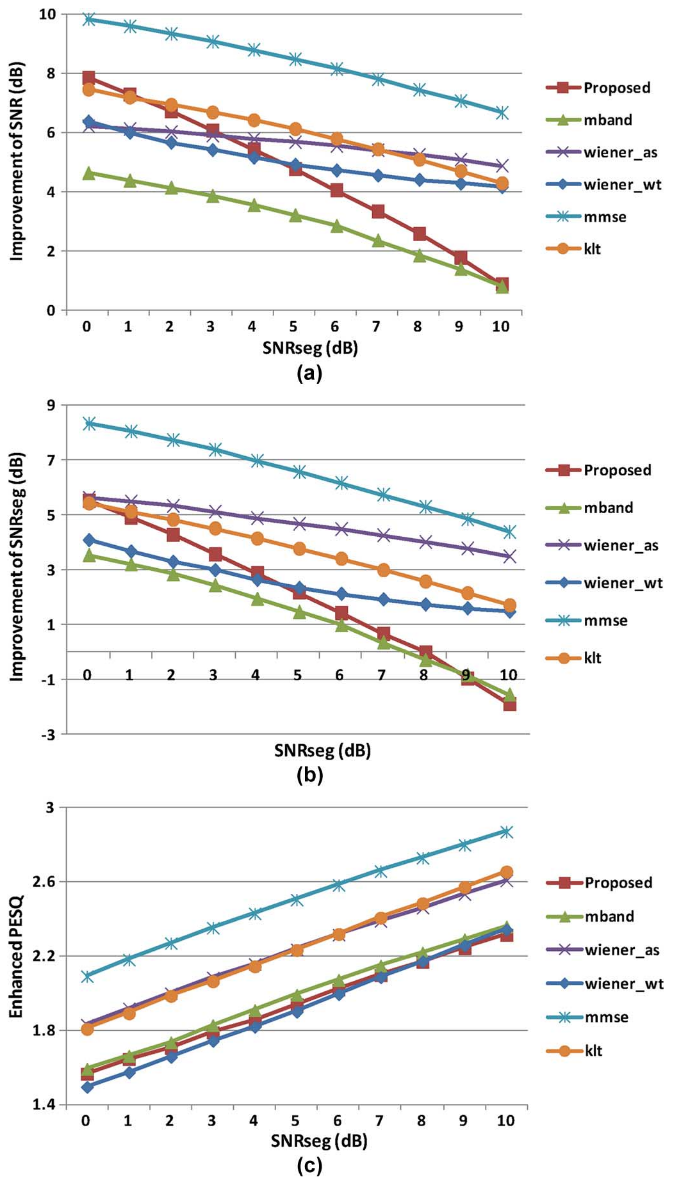

Fig. 11. Comparisons of the SNR, $\mathrm{SNR}_{\mathrm{seg}}$ and PESQ improvements in stationary noise environments (a) SNR improvement. (b) $\mathrm{SNR}_{\mathrm{seg}}$ improvement. (c) Enhanced PESQ.

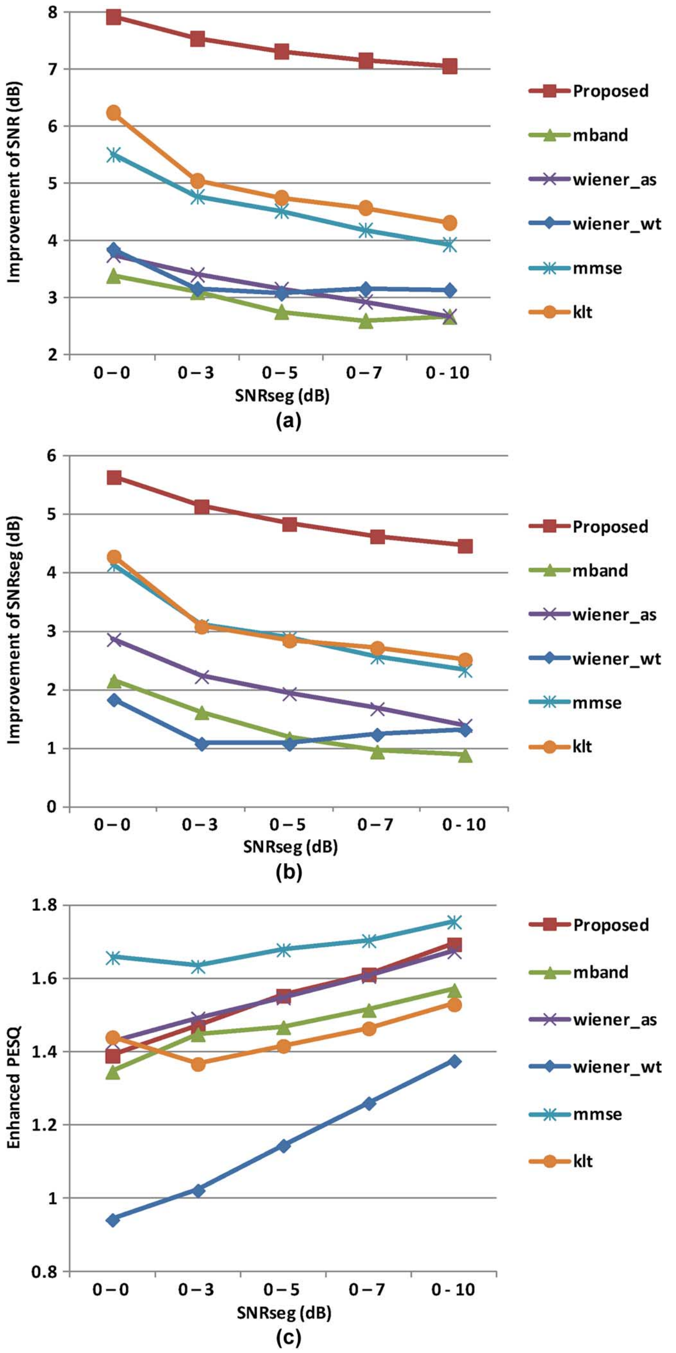

Fig. 12. Comparisons of the SNR, $\mathrm{SNR}_{\mathrm{seg}}$ and PESQ improvements in non-stationary noise environments. The 0-3 in x-axis means the background noise changes dramatically between $\mathrm{SNR}_{\mathrm{seg}}$ 0 dB and $\mathrm{SNR}_{\mathrm{seg}}$ 3 dB, and so on. (a) SNR improvement. (b) $\mathrm{SNR}_{\mathrm{seg}}$ improvement. (c) Enhanced PESQ.

improvement of mmse method is about 2 to 3 dB better than that of our proposed algorithm, its computational complexity is very high as will be shown in Table IV. Also, at speech dominated subbands, noise will be masked by speech due to the masking effect. So our algorithm does not process these subbands to avoid musical effect so SNR improvement will be not that high. The performance indices of the proposed algorithm in the three types of non-stationary noise environments are the same with that in stationary noise environments. However, the performance indices of all other algorithms used in Fig. 11 are seriously degraded because of inaccurate noise estimation due to dramatic background noise variation. The average improvement (enhanced output PESQ—original input PESQ) of PESQ is 0.210/0.216 in high $\mathrm{SNR}_{\mathrm{seg}}$ (5 dB $\sim$ 10 dB/0-5 dB $\sim$ 0-10 dB) stationary/non-stationary noise environment where the patients usually suffer from the speech distortion in this environment. In the non-stationary noise environment, the improvement of PESQ of the proposed algorithm is only a little less than mmse and is better than other algorithms. The good performance both in stationary and non-stationary noise environments and low computational complexity make the proposed algorithm suitable for HA systems.

B. Complexity Comparison

Low power consumption is the most important criterion in HA system implementation. Because low computational com-

TABLE IV
THE COMPARISON OF COMPUTATIONAL COMPLEXITY OF EACH SAMPLE BASED ON THE NUMBER OF MULTIPLICATION, DIVISION, LOGARITHM AND ADDITION

Operation	Proposed Algorithm (Original)	Proposed Algorithm (Implement)	specsub [1]	mband [2]	mmse [3]
Mul.	7.38	2.00	11.25	19.50	21.00
Div.	0.28	0	1.50	4.50	12.00
Log.	0	0	0	1.50	1.50
Add.	14.85	22.78	6.75	13.50	13.50

plexity during speech procession or speech process means low power consumption, the computational complexity comparisons of the proposed algorithm and several selected algorithms in HA system implementation are shown in Table IV. The specsub means the full band spectral subtraction algorithm [1]. The comparison is based on the average usage of multiplication (Mul.), division (Div.), logarithm (Log.) and addition (Add.) per sample in each algorithm. The proposed algorithm (Original) means the original neuromorphic pitch based noise reduction algorithm and the proposed algorithm (Implement) means the proposed algorithm with those low-power architecture designs mentioned at Section III-D. As shown in Table IV, the usage of multiplication of the proposed algorithm (Original) is about 35% to 65% of the other algorithms while the usage of division is about 2% to 18% of the other algorithms. Furthermore, the computational complexity of the proposed algorithm (Implement) is further reduced by using the proposed low power architecture.

V. HARDWARE IMPLEMENTATION RESULT

The proposed NR algorithm with low power architecture design is implemented with the AFC, AFB, WDRC and SFB block of HA system in Fig. 1. The implementation of the proposed system uses cell based flow and 65 nm high $\mathrm{V_T}$ (HVT) CMOS cell library. In addition, the gated clock and operand isolation techniques are applied in order to reduce the power consumption. Fig. 13 shows the layout and die photo of the HA system. Table V is the measurement data of implementation results and power analysis of the proposed NR design. The clock rate of the proposed NR block is 2.5 MHz and the data rate is 24 kHz. The number of cycle count of the proposed architecture is 389 which means the hardware latency incurred for the proposed algorithm is about 0.156 ms. The memory usage of the proposed NR block is 12 Kbits. All the memory used in this chip is implemented by register cell. Finally, the total power consumption of the proposed NR block is 47.74 μW at 0.5 V supply voltage.

VI. CONCLUSION

In this paper, a low computational complexity hardware-oriented neuromorphic pitch based noise reduction algorithm and hardware implementation for monosyllable hearing aid system applications are proposed. The proposed NR algorithm, suitable in ANSI S1.11 or quasi ANSI filter bank architecture, consists of a pitch-based VAD to detect speech duration and a neuromorphic noise attenuator to reduce the background noise. In stationary noise environment, the proposed NR algorithm can improve $\mathrm{SNR_{seg}}$ by 4.238 dB in low $\mathrm{SNR_{seg}}$ environment and improve PESQ by 0.210 in high $\mathrm{SNR_{seg}}$ environment in average. In non-stationary noise environment, the average improvement of $\mathrm{SNR_{seg}}$ is 4.943 dB and the average improvement of PESQ is 0.216. The most important advantage is that unlike multiband (mband) spectral subtraction [2], wiener filter (wiener_as, wiener_wt) [26], [27], minimum mean square error (mmse) [3] and subspace (klt) [5] algorithms, the performance of the proposed neuromorphic pitch based NR algorithm is not degraded in non-stationary noise environment. In addition, the computational complexity of the proposed NR algorithm with the low power architecture can save 90% complexity than the other compared NR algorithms. Therefore, the proposed neuromorphic pitch based noise reduction algorithm is suitable in monosyllable HA system application. Chip implementation shows the proposed NR algorithm consumed 47.74 μW at 0.5 V supply voltage.

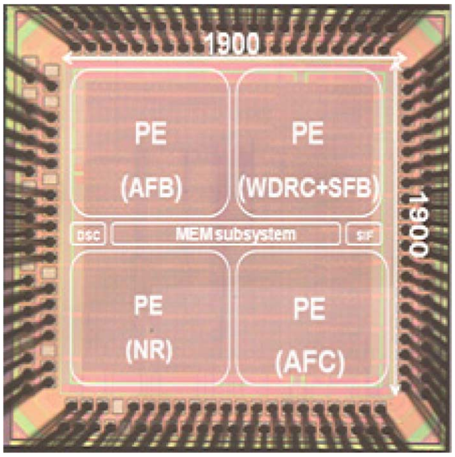

Fig. 13. The die photo of the HA system.

TABLE V
THE MEASUREMENT DATA OF IMPLEMENTATION RESULT OF THE PROPOSED NOSIE REUCTION (NR) BLOCK

Process	65nm CMOS HVT
Clock Rate	2.5 MHz
Data Sampling Rate	24 KHz
Supply Voltage	0.5V
NR Core Area	830 μm *710 μm
Power Consumption	47.74 μW

REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1979, vol. 4, pp. 208–211.

[2] S. Kamath and P. C. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, 2002, vol. 4, pp. 4164–4167.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *Proc. IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[4] P. C. Loizou, “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, 2005.

[5] H. Yi and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, 2003.

[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1979, vol. 4, pp. 208–211.

[7] J. C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Proc. Conf. Eurospeech*, 1991, pp. 1371–1374.

[8] C. W. Wei, C. C. Tsai, T. S. Chang, and S. J. Jou, "Perceptual multiband spectral subtraction for noise reduction in hearing aids," in *Proc. IEEE Int. Conf. APCCAS*, May 2010, pp. 692–695.

[9] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust Voice Activity Detection Using Long-Term Signal Variability," *Proc. IEEE Trans. Audio, Speech, Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[10] M. A. Stone and B. C. J. Moore, "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production," *J. Ear Hearing*, vol. 23, no. 4, pp. 325–338, 2002.

[11] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement," *Signal Processing, IET*, vol. 6, no. 1, pp. 54–63, 2012.

[12] Y. T. Kuo, T. J. Lin, Y. T. Li, and C. W. Liu, "Design & implementation of low-power ANSI S1.11 filter bank for digital hearing aids," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1684–1696, Jul. 2010.

[13] C. W. Liu, K. C. Chang, M. H. Chuang, and C. H. Lin, "Design and implementation of 18-band Quasi-ANSI S1.11 1/3-octave filter bank for digital hearing aids," in *Proc. IEEE Int. Symp. VLSI Design, Automation, Test*, April 2012, pp. 1–4.

[14] *Specification for Octave-Band and Fractional-Octave-Band Analog and Digital Filters*, ANSI S1.11-2004, Standards Secretariat Acoustical Society of America, Feb. 2004.

[15] C. W. Wei, Y. T. Kuo, K. C. Chang, C. C. Tsai, J. Y. Lin, Y. FanJiang, M. H. Tu, C. W. Liu, T. S. Chang, and S. J. Jou, "A low-power Mandarin-specific hearing aid chip," in *Proc. IEEE Asian Solid-State Circuit Conference*, Nov. 2010, pp. 1–4.

[16] Y. T. Kuo, T. J. Lin, W. H. Chang, Y. T. Liu, and C. W. Liu, "Complexity-effective auditory compensation for digital hearing aids," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2008, pp. 1472–1475.

[17] K. C. Chang, Y. T. Kuo, T. J. Lin, and C. W. Liu, "Complexity-effective dynamic range compression for digital hearing aids," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 2378–2381.

[18] D. F. Rosenthal and H. G. Okuno*, Computational Auditory Scene Analysis*. Mahwah, NJ, USA: Lawrence Erlbaum, 1998.

[19] , D. Wang and G. J. Brown, Eds.*, Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. New York, NY, USA: Wiley-IEEE Press, 2006.

[20] J. Chalupper, "Aural Exciter and Loudness Maximizer: What's Psychoacoustic about" Psychoacoustic Processors?," in *Proc. 109th AES Convention*, 2000, pp. 1472–1475.

[21] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2236–2252, 2003.

[22] S. Mukhopadhyay and G. C. Ray, "A new interpretation of nonlinear energy operator and its efficacy in spike detection," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 2, pp. 180–187, Feb. 1998.

[23] P. C. Loizou*, Speech Enhancement, Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[24] A. Varga, H. J. M. Steenneken, M. Tomlinson, and D. Jones, 1992, NOISEX-92 [Online]. Available: http://spib.rice.edu/spib/select_noise.html

[25] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, 2001, pp. 749–752.

[26] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, 1996, vol. 2, no. 7, pp. 629–632.

[27] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 59–67, 2004.

[28] H. Dillon*, Hearing Aids*. New York, NY, USA: Boomerang Press, 2001.

Yu-Jui Chen received the B.S. and M.S. degrees in electronic engineering from National Chiao Tung University (NCTU,) Hsinchu, Taiwan, in 2009 and 2011, respectively.

His research interests include digital IC design and speech processing.

Cheng-Wen Wei received the B.S. and M.S. degrees in electrical engineering from the Yuan Ze University, Taiwan, in 1998 and 2000 and received the Ph.D. degree in electronic engineering from National Chiao Tung University (NCTU,) Hsinchu, Taiwan in 2012.

From 2000 to 2006, he was an engineer and worked on Delta Sigma data convertor, speech signal processing and VLSI design, with the Product Development Division/Digital Circuit Design Dept. (DCD,) Elan Microelectronics Corporation (EMC,) Hsinchu, Taiwan. Since 2006, he has been a consultant with the DCD, EMC. His research interests include digital signal processing, speech processing, data conversion and low power VLSI design.

Yi FanChiang received the B.S. degree in electrical engineering from the National Central University, Taoyuan, Taiwan, in 2006 and M.S. degree from National Tsing Hua University, Hsinchu, Taiwan in 2008. He is currently working toward the Ph.D. degree in electronic engineering from National Chiao Tung University (NCTU,) Hsinchu, Taiwan since 2008.

His research interests include digital signal processing, speech processing, bio-electronic circuit and blind sources separation.

Yi-Le Meng received the B.S. and M.S. degrees in electronic engineering from National Chiao Tung University (NCTU,) Hsinchu, Taiwan, in 2011.

His research interests include low-power digital IC design and speech processing.

Yi-Cheng Huang received the B.S. degree in electronic engineering from National Chiao Tung University (NCTU,) Hsinchu, Taiwan, in 2011. He is currently working toward the M.S. degree in electronic engineering from National Chiao Tung University (NCTU,) Hsinchu, Taiwan since 2011.

His research interests include digital signal processing, speech processing, bio-electronic circuit and low power VLSI design.

Shyh-Jye Jou (S'86–M'90–SM'97) received the B. S. degree in electrical engineering from National Chen Kung University in 1982, and the M.S. and Ph.D. degrees in electronics from National Chiao Tung University in 1984 and 1988, respectively.

He joined Electrical Engineering Department of National Central University, Chung-Li, Taiwan, from 1990 to 2004 and became a Professor in 1997. Since 2004, he has been a Professor of National Chiao Tung University and became the Chairman of Department of Electronics Engineering from 2006 to 2009. From 2011 he becomes the Vice President, Office of International Affair, National Chiao Tung University. He was a visiting research Professor in the Coordinated Science Laboratory at University of Illinois, Urbana-Champaign during 1993–1994 and 2010 academic years. In the summer of 2001, he was a visiting research consultant in the Communication Circuits and Systems Research Laboratory of Agere Systems, USA.

Dr. Jou received Outstanding Engineering Professor Award, Chinese Institute of Engineers at 2011. He was the Guest Editor, IEEE JOURNAL OF SOLID STATE CIRCUITS, Nov. 2008. He served as the Conference Chair of IEEE International Symp. on VLSI Design, Automation and Test (VLSI-DAT) and International Workshop on Memory Technology, Design, and Testing. He also served as Technical Program Chair or Co-Chair in IEEE VLSI-DAT, International IEEE Asian Solid-State Circuit Conference, IEEE Biomedical Circuits and Systems, and other international conferences. He has published more than 100 IEEE journal and conference papers. His research interests include design and analysis of high speed, low power mixed-signal integrated circuits, communication and Bio-Electronics integrated circuits and systems.