Technology and Health Care 26 (2018) S281–S289
DOI 10.3233/THC-174615
IOS Press

Noise reduction algorithm with the soft thresholding based on the Shannon entropy and bone-conduction speech cross-correlation bands

Sung Dae Na[a], Qun Wei[b], Ki Woong Seong[c], Jin Ho Cho[d] and Myoung Nam Kim[e,*]
[a]*Department of Medical and Biological Engineering, Graduate School, Kyungpook National University, Daegu, Korea*
[b]*Department of Biomedical Engineering, School of Medicine, Keimyung University, Daegu, Korea*
[c]*Department of Biomedical Engineering, Kyungpook National University Hospital, Daegu, Korea*
[d]*School of Electronics Engineering, College of IT Engineering, Kyungpook National University, Daegu, Korea*
[e]*Department of Biomedical Engineering, School of Medicine, Kyungpook National University, Daegu, Korea*

Abstract.
BACKGROUND: The conventional methods of speech enhancement, noise reduction, and voice activity detection are based on the suppression of noise or non-speech components of the target air-conduction signals. However, air-conduced speech is hard to differentiate from babble or white noise signals.
OBJECTIVE: To overcome this problem, the proposed algorithm uses the bone-conduction speech signals and soft thresholding based on the Shannon entropy principle and cross-correlation of air- and bone-conduction signals.
METHODS: A new algorithm for speech detection and noise reduction is proposed, which makes use of the Shannon entropy principle and cross-correlation with the bone-conduction speech signals to threshold the wavelet packet coefficients of the noisy speech.
RESULTS: The proposed method can be get efficient result by objective quality measure that are PESQ, RMSE, Correlation, SNR.
CONCLUSION: Each threshold is generated by the entropy and cross-correlation approaches in the decomposed bands using the wavelet packet decomposition. As a result, the noise is reduced by the proposed method using the MATLAB simulation. To verify the method feasibility, we compared the air- and bone-conduction speech signals and their spectra by the proposed method. As a result, high performance of the proposed method is confirmed, which makes it quite instrumental to future applications in communication devices, noisy environment, construction, and military operations.

Keywords: Noise reduction, bone conduction, Shannon entropy, speech

1. Introduction

Nowadays, there is a large demand for speech signal-processing applications for speech enhancement

*Corresponding author: Myoung Nam Kim, Department of Biomedical Engineering, School of Medicine, Kyungpook National University, Daegu 41944, Korea. E-mail: kimm@knu.ac.kr.

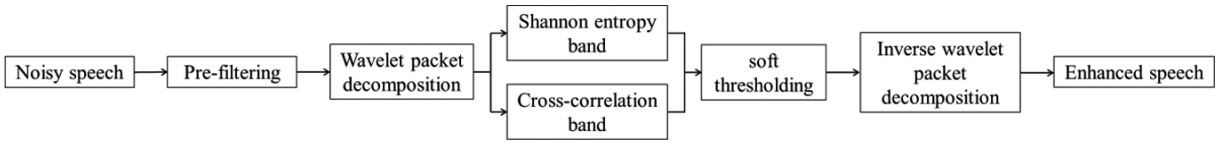

Fig. 1. The proposed algorithm flowchart.

in the communication devices. On the one hand, the hearing loss problems of older people, which refer to the aging society challenges, are classified into conductive, sensorineural, and mixed hearing loss cases that require different solutions and hearing aid equipment [1]. On the other hand, various critical situations, construction works in the noisy environment, and military operations require the robust enhancement of noisy speech, in order to deliver the correct information and avoid the disastrous consequences of lost signals, possible misunderstanding, or "some wires getting crossed". The traditional speech enhancement, noise reduction, and voice activity detection algorithms are based on linear processing techniques and general air-conduction speech signals [2]. Although these methods suppress the background noises, they are prone to distort the target speech signal and tend to introduce perceptually annoying residual noises [3–5]. Speech enhancement methods can bedivided into several categories based on their domains of operation, namely time domain, frequency domain, and time-frequency domain. Time domain methods include the subspace approach, frequency domain methods include short-time Fourier transform-based spectral subtraction, minimum mean square error estimator and Wiener filtering, while time frequency-domain methods involve the employment of the family of wavelets. Recently, wavelet transform has been widely applied to the signal and image de-noising, compression, detection, and pattern recognition problems. Thus, Tahsina proposed the wavelet packet coefficients method, which involves the semi-soft thresholding and Teager energy operator (TEO). The latter is applied to compute a threshold value that is used to threshold the wavelet packet coefficients of the noisy speech. However, it is not always possible to separate the components corresponding to the target signal from those of noise by a simple thresholding in the babble and white noise environment [6–8]. Recently, the improved algorithms using a bone-conducted microphone have been proposed, e.g., by Li. Bone conduction is the conduction of sound to the inner ear through the bones of the skull, which can be used by individuals with normal or impaired hearing. The bone-conducted microphone tunes-off the outside noise due to the different principles of sound transmission via air and bone vibrations. However, the method is less effective dealing with the residual and musical noises, and frequently fails to satisfy the speech signal continuity and prevent the speech signal losses [9,10].

In this paper, a new algorithm for speech detection and noise reduction is proposed, which makes use of the Shannon entropy principle and cross-correlation with the bone-conduction speech signals to threshold the wavelet packet coefficients of the noisy speech. The proposed speech enhancement algorithm exhibits a good performance of noise reduction and extraction of the speech signal information from the simulated noisy speech patterns with babble and white noises.

2. Method

2.1. The proposed algorithm flowchart and experimental model

The proposed algorithm flowchart is depicted in Fig. 1. Firstly, the noisy speech signal pre-filtering is conducted according to different signal transmit principles of air- and bone-conduction: the former

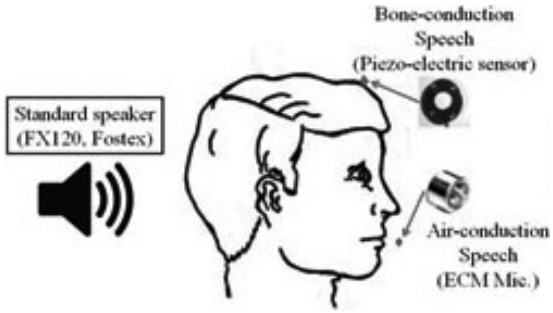

Fig. 2. The experimental model for speech enhancement in the robust environment.

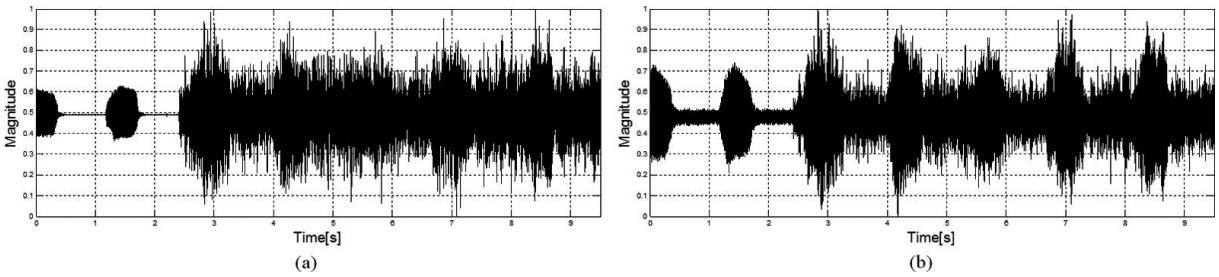

Fig. 3. Air-conduction (a) and bone-conduction (b) speech signals.

involves propagation of sound waves in the air, while the latter envisages the conduction of sound to the inner ear through the vibrations of head skin and skull bones. Therefore, a 100 Hz high-pass filter is applied to the noisy speech signals to calibrate the difference signals and remove the unnecessary noises. Secondly, each speech signal is decomposed using the wavelet packet, as a simple and powerful de-noising tool. Consequently, the filtered signals are subjected to the wavelet packet decomposition to extract the speech and noise components and assess their duration. The Shannon entropy and cross-correlation thresholds are calculated for the noise reduction in each band by the proposed method, which will be discussed in more detail in the next section. Finally, a soft thresholding method is applied to the speech bands using the calculated threshold values for the speech enhancement and noise reduction, and the inverse wavelet packet decomposition is used to provide the enhanced speech output.

Figure 2 depicts the experimental model. To acquire the speech and noise signals in a robust environment, the electric condenser microphone (ECM) and piezo-electric sensor are used simultaneously in the experimental model. The former is placed close to the speaking person's mouth to acquire the air-conduction speech signals, and the latter is fixed at the top of the person's head, which is the most suitable position for deriving the bone-conduction signals [11]. To simulate a noisy environment, a standard Fostex FX120 speaker, which is a 125 mm cone type full-range speaker with 100 dB SPL, is used to produce noises. The speaker cone is made of KENAF fiber composite and is suspended with a rubber surround. The vowel 'a' is pronounced several times at about 70–75 dB SPL for the test repeatability.

The conventional methods of speech enhancement have problems with differentiation of the speech signals from the babble noises, because they have very similar features. Therefore, we proposed the new algorithm using the bone-conduction speech signal acquired at the top of the head. Moreover, the signals obtained via the ECM and piezoelectric sensor were simultaneously acquired during the experiment.

The experiment was conducted in a chamber with a background noise of approximately 30–35 dB SPL. The speech signals were generated repeatedly, while a Fast Ultra Track 8R (M-Audio) audio-band

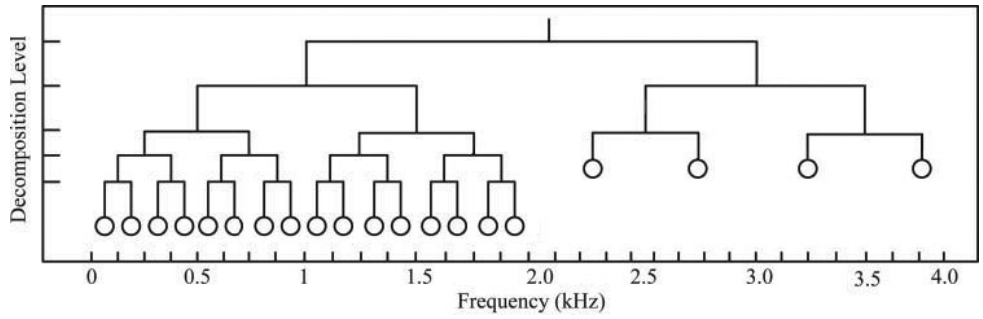

Fig. 4. Band structure of the modified wavelet packet decomposition (MWPD).

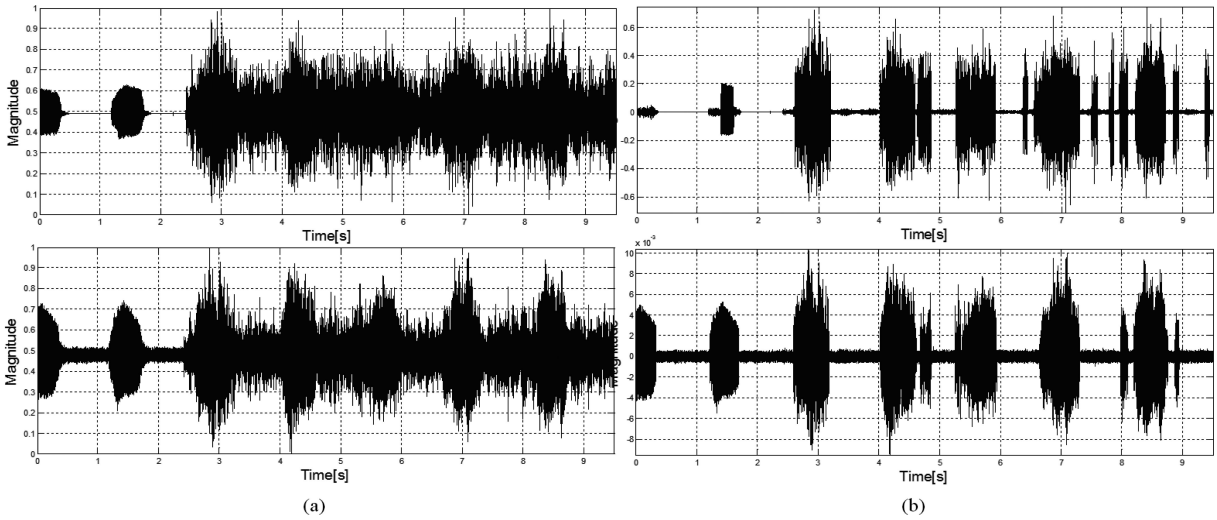

Fig. 5. (a) Original signals obtained by air-conduction (above) and bone-conduction (below); b) Results of the proposed method application to air-conduction (above) and bone-conduction (below) signals.

measurement and analysis device was applied to the noisy speech signals acquired via the ECM and piezo-sensor. The measured signals are plotted in Fig. 3.

2.2. Proposed speech enhancement using entropy and correlation of each band

The wavelet packet decomposition is widely used in speech signal processing since it is simple and powerful tool, which can resolve frequency components within a small time window of a speech signal. The bone conduction speech signals are mainly distributed in the low-frequency band and are relatively less affected by strong external environment, as compared to the conventional air conduction voice signals. However, to extract noise information, such as white and babble noises, which are distributed in all bands or are similar to voice signals, a time-frequency domain method, which allows one to verify the information of each frequency band and time domain, seems more lucrative than time domain and frequency domain ones. Due to the bone-conduction speech signal characteristics, the wavelet packet decomposition can easily compare and extract its features.

A noisy speech signal $y(n)$ can be described as:

$$y(n) = s(n) + v(n) \tag{1}$$

where $s(n)$ is clean speech and $v(n)$ is background noise in nth frame. Wavelet packet decomposition is applied to the noisy signal using the wavelet packet transform into time-frequency wavelet coefficients of multiple sub-bands. The decomposition of multiple sub-bands is readily applicable to the bone-conduction speech [10].

In this paper, wavelet packet decomposition was modified to reduce the noise and detect the speech bands based on the Daubechies D6 wavelet, which encodes three polynomials, i.e., constant, linear, and quadratic signal components. The speech signal is decomposed into 20 sub-bands with the wavelet coefficient $w_{j,m}(k)$ using the modified wavelet packet decomposition (MWPD). To correlate the speech signals with the bone conduction ones, the wavelet decomposition is conducted at the low-frequency band. Figure 4 shows the designed decomposition bands. The proposed method provides the signal decomposition into the sub-bands that are mainly distributed in the low-frequency band, so that the number of sub-bands in the high-frequency range is smaller than that in the low-frequency one. Also, the numbers of sub-bands are modified based on the bone-conduction speech that is generated by the vibration propagating from the vocal cord to the temporal bone. Consequently, the decomposition yields 20 sub-bands at the fourth level of the tree (see Fig. 4), which provides the sufficient resolution of each low-frequency band. In other words, $w_{j,m}(k)$ represents the jth level, kth wavelet coefficient of the mth sub-band in the MWPD, where $j = 2, 3, 4$, $m = 1, 2, \ldots, 20$, and $k = 1, 2, \ldots, N/2j$. The following matrix describes the modified $w_{j,m}(k)$:

$$\Psi_m(t) = \begin{bmatrix} \psi_1(t) \\ \psi_2(t) \\ \vdots \\ \psi_4(t) \end{bmatrix} = \begin{bmatrix} \psi_1(1) & \psi_1(2) & \cdots & \psi_1(t) \\ \psi_2(1) & \psi_2(2) & & \vdots \\ \vdots & \cdots & \ddots & \vdots \\ \psi_{20}(1) & \psi_{20}(2) & \cdots & \psi_{20}(t) \end{bmatrix} \tag{2}$$

where $\Psi_m(t)$ is the composed signal of the mth sub-band at the particular time t, and $\Psi_m(t)$ is a matrix of the wavelet coefficient information in time and wavelet bands. Figure 4 depicts the band structure of the modified wavelet packet decomposition (MWPD) for $\Psi_m(t)$ corresponding to the bone-conduction speech signal. Insofar as frequency features of the bone conduction signals are generated in the low-frequency band, the bands of $\Psi_m(t)$ are also distributed in the low-frequency band.

In the proposed method, for the implementation of WP decomposition, the Matlab wavelet toolbox is used, where in order to obtain the optimal decomposition, the Shannon entropy criterion is employed. The proposed algorithm uses two thresholds assessed via the entropy domain and cross-correlation domain approaches. The $\psi_1(1)$ signal is very similar to the original signal. In other words, the speech and noise signal are composed using the information on $\psi_1(1)$. The entropy and cross-correlation criteria are used to estimate the thresholds of speech and noise information, respectively. The entropy and cross-correlation are calculated for each band, while the correlation threshold is assessed via $\psi_1(1)$. The speech and noise signals are combined based on the extracted values, whereas the speech information exhibits a higher entropy and closer correlation with $\psi_1(1)$.

Firstly, the speech and noise signals can be assessed using the energy-based approach, which can be reduced to the following equation:

$$\begin{cases} \Psi_S(t) = \sum\limits_{n=1}^{m} \Psi_m(t), & \Psi_m(t) > \Psi_{th}(t) \\ \Psi_N(t) = \sum\limits_{n=1}^{m} \Psi_m(t), & \Psi_m(t) < \Psi_{th}(t) \text{ where, } \Psi_{th}(t) = \dfrac{1}{m} \sum\limits_{n=1}^{m} |\Psi_m(t)|^2 \end{cases} \tag{3}$$

where $\Psi_S(t)$ and $\Psi_N(t)$ are speech and noise signals, respectively, while $\Psi_{th}(t)$ is threshold value, which is used to differentiate speech and noise signals in $\Psi_m(t)$. In other words, the speech signal

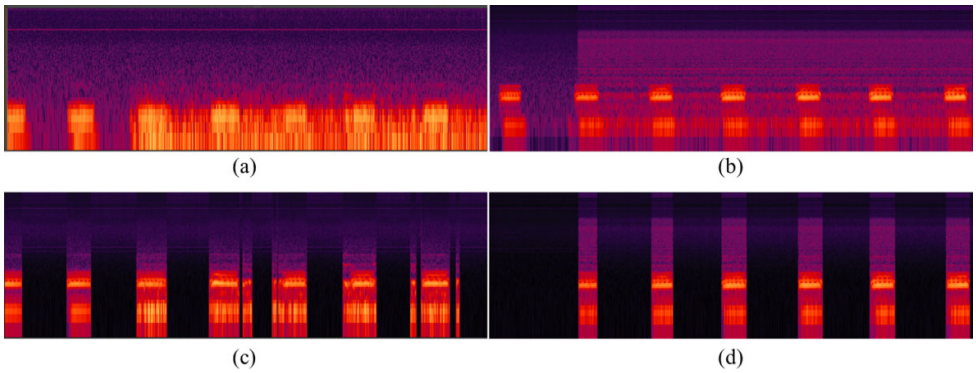

Fig. 6. Spectra of signals: (a) original speech with babble noise, (b) original speech with white noise, (c) result of the proposed method application to speech with babble noise, (d) result of the proposed method application to speech with white noise.

and noise signal can be assessed using Eq. (3). Nevertheless, the combined signal still contains the noise information, for which reduction the threshold is applied to the signal according to the proposed method.

The threshold value can be derived from the following equation via the entropy of each band that is subjected to the wavelet packet decomposition:

$$EG(t) = \begin{cases} 1, & \frac{1}{N}\sum_{n=1}^{k} \Psi_e(t) \log \Psi_e(t) > \Psi_e(t) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $EG(t)$ is the entropy band threshold and $\Psi_e(t)$ is the combined signal via the entropy criterion. The speech or voice duration is detected by $EG(t)$ because the speech information has higher entropy value than noise, and the bone-conduction signal is less affected/distorted by the external noise. However, this procedure is not sufficient to retrieve the speech information. To overcome this problem, the proposed method involves another thresholding procedure. The second threshold can be estimated using the following equation:

$$CG(t) = \begin{cases} 1, & \sqrt{\frac{\sum_{k=1}^{n} \Psi_{ck}^2}{n} - m^2} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $CG(t)$ is the correlation band threshold, which is derived by the cross-correlation using $\Psi_1(1)$ that is similar to the original signal after decomposition, while $\Psi_c(t)$ is the combined signal by cross-correlation.

The final threshold is generated using $EG(t)$ and $CG(t)$, which are also applied for the voice detection. The speech information can be extracted via the thresholds, insofar as speech information is distributed in high-entropy and high-correlation bands.

3. Results and discussion

The processed data are shown in Fig. 5, wherer the original speech signals are plotted in Fig. 5a. The bone-conduction speech signal is acquired by the piezo-electric sensor. Therefore, the bottom part of Fig. 5a provides the power-noise baseline. Also, the signal with vowel ‘a’ is shown in Fig. 5a as repeatedly pronounced until 2 s, while the noisy signal containing vowel ‘a’, babble, and white noise

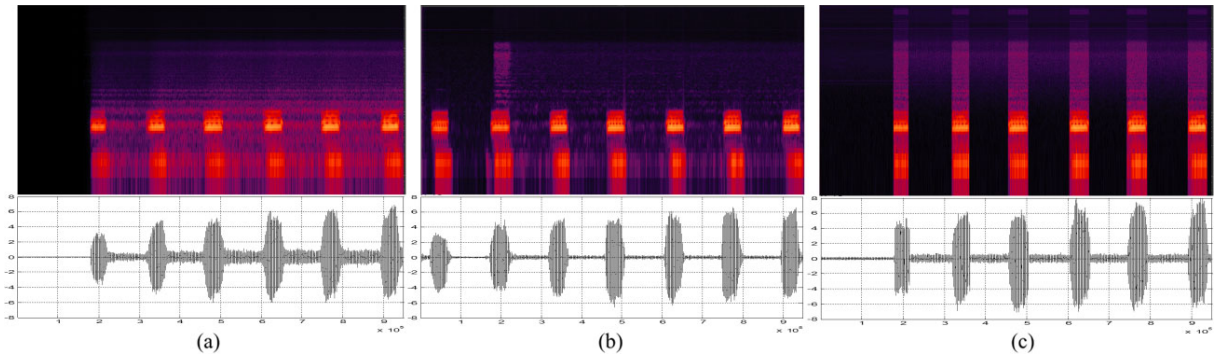

Fig. 7. Results of signal and spectrum: (a) adaptive filter, (b) spectral subtraction, (c) proposed method.

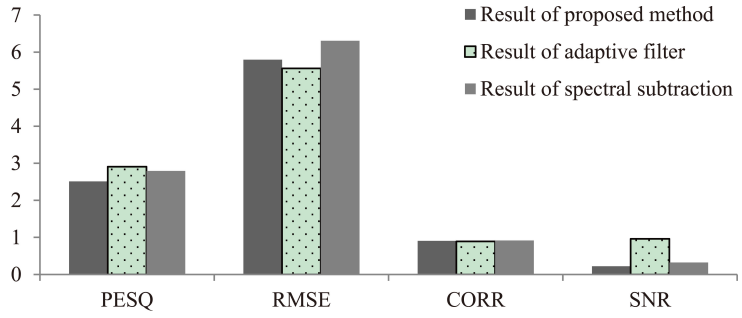

Fig. 8. Performance of comparison between conventional method and proposed method.

appears starting from 2.5 s. Figure 5b shows the result of proposed method application to the air- and bone-conduction speech signals. The conventional air-conduction signal is a useful and simple speech signal, but it can be easily distorted by other speech signal and environmental noise. The major frequency of bone-conduction signal falls into the low-frequency band, and its generation principle is different from the air-conduction signal. In other words, the bone-conduction signals are less sensitive to the effect of other speech and environmental noise than the air-conduction ones. Therefore, we applied the proposed method to air- and bone-conduction signals to verify its efficiency. Figure 5b shows the result of the proposed method application to the bone-conduction signals. The noise is efficiently removed using the entropy and cross-correlation thresholds. However, vowels of 4th and 7th series can contain some noises because of low SNR or a noisy environment. Finally, in the upper part of Fig. 5b the vowel 'a' signals of 1th and 2th series are removed due to the white noise presence, although the signals are decomposed by the wavelet packet decomposition. Moreover, the residual noises remain at 6th and 7th series.

Figure 6 shows the spectra of signals. The proposed method exhibits a good de-noising performance as applied to babble and white noise. However, Fig. 6c reveals a problem with the fourth and last speech series because of low SNR. Figure 6d shows the result of white noise suppression, but the first series is removed because white noise is distributed along the whole frequency band.

In addition, efficient of proposed method is compared by conventional methods which are adaptive filter and spectral subtraction. The Fig. 7 shows the comparison between proposed method and others. The result of adaptive filter in Fig. 7a can be confirm the residual noise in spectrum. And the spectral subtraction method is good performance in time domain at Fig. 7b. However, residual noise and distortion of signal can be show in spectrum.

In order to evaluate the algorithm, we conducted the evaluation by objective quality measure [?]. These parameters can be expressing the efficiency of algorithms. The PESQ (perceptual evaluation of speech quality) is a family of standards comprising a test methodology for automated assessment of the speech quality as experienced by a user of a telephony system and can be shown the quality of algorithm between original signal and after processing signal. And RMSE (root mean square error) and correlation can be shown the accuracy ratio between original signal and results. The proposed method enough to take the efficiency likes conventional method. However, in order to improve the efficiency of result, the algorithms have to research more.

4. Conclusion

This paper proposes a noise reduction and voice detection method that applies two thresholds and uses the bone-conduction speech signal with the babble noise. Each threshold is generated by the entropy and cross-correlation approaches in the decomposed bands using the wavelet packet decomposition. As a result, the noise is reduced by the proposed method using the MATLAB simulation. To verify the method feasibility, we compared the air- and bone-conduction speech signals and their spectra by the proposed method. As a result, high performance of the proposed method is confirmed, which makes it quite instrumental to future applications in communication devices, noisy environment, construction, and military operations.

Acknowledgments

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) and the Ministry of Education (Nos. 2015R1A2A2A03006113, 2016R1A 2A1A05005413, 2017M3A9E2065284, and 2017R1D1A1B03031388)

Conflict of interest

None to report.

References

[1] Kim HP, Han JH, Kwon SY, Lee SM, Kim DW, Hong SH, Kim IY, Kim SI. Sensitivity enhancement of speech perception in noise by sound training: Hearing loss simulation study. Biomedical Engineering Letters 2011; 1(2): 137-142.
[2] Han JH, Yook SH, Nam KW, Lee SM, Kim DW, Hong SH, Jang DP, Kim IY. Comparative evaluation of voice activity detectors in single microphone noise reduction algorithms. Biomedical Engineering Letters 2012; 2(4): 255-264.
[3] Lei SF, Tung YK. Speech enhancement for nonstationary noises by wavelet packet transform and adaptive noise estimation. In Intelligent Signal Processing and Communication Systems, Proceeding of International Symposium on IEEE 2005; 41-44.
[4] Chung K. Challenges and recent developments in hearing aids, Part I, Speech understanding in noise, microphone technologies and noise reduction algorithms. Trends Amplif 2004; 8: 83-124.
[5] Levitt H. Noise reduction in hearing aids: A review. J Rehabil Res Dev 2001; 38: 111-21.
[6] Tahsina FS, Celia S. A semisoft thresholding method based on teager energy operation on wavelet packet coefficients for enhancing noisy speech. Journal on Audio, Speech, and Music Processing 2013.
[7] Bhoura M, Rouat J. Wavelet speech enhancement based on the teager energy operator. Signal Processing Letters IEEE 2001; 8(1): 10-12.

[8] Gao HY, Bruce AG. Waveshrinkage with semisoft shrinkage. Statsci Divition of Mathsoft Inc 1995.
[9] Mingzi L, Israel C, Saman M. Multisensory speech enhancement in noisy environments using bone conducted and air conducted microphones. Proceeding of IEEE China Summit and International Conference on Signal and Information Processing 2014.
[10] Li Y, Wang D. On the optimality of idel binary time-frequency masks. Speech Commun 2009; 51(30): 230-239.
[11] Lee JN, Lee GH, Na SD, Ki WS, Cho JH, Kim MN. Noise cancellation algorithm of bone conduction speech signal using feature of noise in separated band. Journal of Korea Multimedia Society 2016; 19(2): 128-137.