ISSN 2007-9737

Post-Processing for the Mask of Computational Auditory Scene Analysis in Monaural Speech Segregation

Wen-Hsing Lai, Cheng-Jia Yang, Siou-Lin Wang

Institute of Computer and Communication Engineering Kaohsiung First,
University of Science and Technology, Taiwan,
China

{lwh, u0151819, u0015901}@nkfust.edu.tw

Abstract. Speech segregation is one of the most difficult tasks in speech processing. This paper uses computational auditory scene analysis, support vector machine classifier, and post-processing on binary mask to separate speech from background noise. Mel-frequency cepstral coefficients and pitch are the two features used for support vector machine classification. Connected Component Labeling, Hole Filling, and Morphology are applied on the resulting binary mask as post-processing. Experimental results show that our method separates speech from background noise effectively.

Keywords. CASA, Connected Component Labeling, SVM.

1 Introduction

Human auditory system can clearly distinguish the noise and speech even in noisy environment, such as baseball fields, construction sites, or factories. But the recognition rate of a speech or speaker recognition system can decline a lot by the influence of background noise. Over the last few decades, many advances have been made in the area of speech segregation/separation, such as Computational Auditory Scene Analysis (CASA) [1], independent component analysis (ICA) [2], blind source separation (BSS) [3,4], etc. CASA comes from the Auditory Scene Analysis (ASA) which Bregman proposed. ASA have a great influence for the later studies [5]. Bregman divided the system into segmentation and grouping stages. The segmentation is to divide input sound into small Time-Frequency units (T-F units) called segment, and grouping is to combine the segments which may come from the same source into a 'group' called stream. Wang [1] used it to simulate human auditory system and solved monaural speech segregation problem. The computational goal of CASA is to obtain an estimated binary mask close to an ideal binary mask. Binary mask can be considered as a T-F unit filter, which pass the target speech and filter out the background noise by setting speech units 1 and noise units 0 [7]. An ideal binary mask, which differentiates target speech and background noise, can be determined by signal-to-noise ratio (SNR). If the SNR is greater than a threshold, it will be labeled as speech; otherwise, it will be labeled as noise. Estimated binary mask can generally be obtained from a classifier.

In this paper, we use cochlear auditory models and inner hair cells model to simulate the human ear of the inner ear auditory characteristics. Next, Mel-frequency cepstral coefficients (MFCCs) and pitch are used as the features of support vector machine (SVM) [8] classifier. Finally, post-processing technique such as Connected Component Labeling [9], Hole Filling [10], and Morphology [11] are applied on the resulting binary mask as post-processing.

Five kinds of noise with different frequency characteristics are used in our experiments, including three kinds of noise used in both training and testing, and two kinds of noise used in testing only. We called them matched and unmatched noise.

Section 2 presents our system configuration, and section 3 describes the post-processing technique used on binary mask. Section 4 shows the experimental results. Conclusions are made in the final section.

2 System Configuration

Our system configuration is as in Figure 1. Gammatone filters [12] are used to model human auditory filters, which are called critical bands. The input is the sound mixture and the output in each channel is divided into overlapping frames. It produces T-F units of the sound mixture. MFCCs and pitch are used as the features of SVM to classify speech units and noise units. Then, we use post-processing technique on binary mask to improve the speech classification performance. The technique includes Connected Component Labeling, Hole Filling, and Morphology. After obtaining a binary mask from SVM classifiers, the segregated speech is resynthesized.

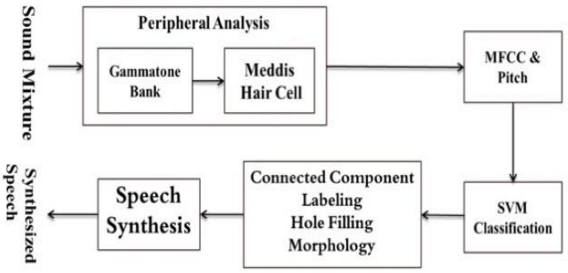

Fig. 1. System configuration

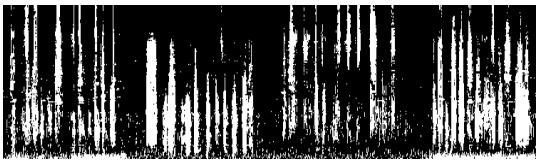

Fig. 2. An example binary mask

3 The Post-Processing on Binary Mask

In this paper, we use post-processing technique on binary mask to improve the speech classification performance. The technique includes Connected Component Labeling, Hole Filling, and Morphology.

3.1 Connected Component Labeling and Hole Filling

The binary mask got from SVM, as an example shown in Figure 2, can be treated as a two dimensional image. The image's height is the number of the channels of Gammatone Bank, and the image's width is the number of the speech frames. The foreground (white blocks) in Figure 2 indicates the speech region, and the background (black blocks) indicates the noise region which should be filtered. We can see many isolated and unconnected white or black blocks in Figure 2. These isolated and unconnected blocks on a binary mask can be considered as the classification error. We, firstly, tried to use Connected Component Labeling and Hole Hilling to fix the problem.

Connected Component Labeling is an algorithm to label the unconnected component in image processing. Commonly used are 4-connected and 8-connected. Those pixels which are connected horizontally or vertically are considered to be the same object in 4-connected,and those pixels which

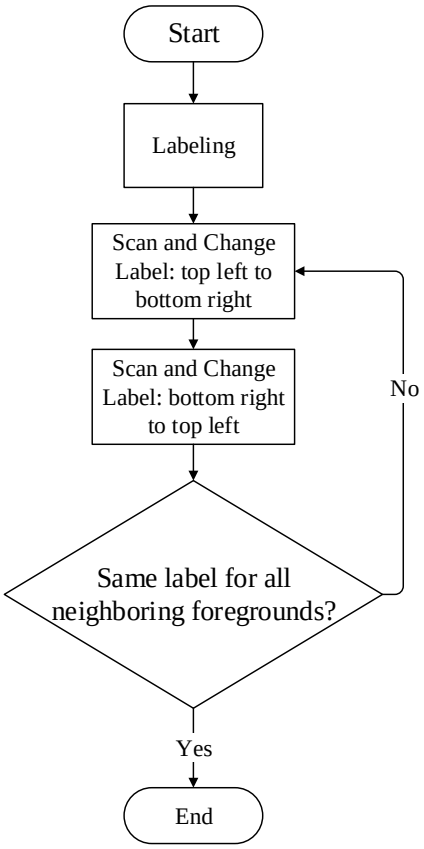

Fig. 3. Connected Component Labeling

are connected horizontally, vertically, or diagonally are considered to be the same object in 8-connected. 4-connected is used in our experiment.

The procedure is as presented in Figure 3. First, label each foreground pixels sequentially. Second, scan and change label from top left to bottom right. If the label of the current pixel is larger than the labels of upper pixel or left pixel, change it to the smallest label number. Again, scan and change label from bottom right to top left. The scan and change procedure will be repeated until all neighboring foregrounds have the same label. At last, we got the area of each connected foreground (speech) objects. Those isolated small area less than 2 points will be reclassified as background (noise).

For the holes on the foreground (speech), we use Hole Filling. Scanning from top left to bottom right, if one background pixel is surrounded by at least 3 pixels in 4 neighbors, we change the background pixel to foreground pixel. Only one scan is done. The binary mask after applying Connected Component Labeling and Hole Filling on Figure 2 is shown in Figure 4.

3.2 Morphology

Morphology is a popular algorithm in image processing to make the contour of objects smooth. It is used on our estimated binary mask to smooth the spectrogram. We applied one time Erosion and Dilation on the mask. Firstly, a foreground pixel is changed to background if it has a background pixel as a 4-neighbor. This procedure is called Erosion. Then, in Dilation, a background pixel is changed to foreground if it has a foreground pixel as a 4- neighbor.

4 Experiments

The clean speech corpus we used in our experiments is extracted from MAT-160 database recorded by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). It is divided into training set and testing set. 30 sentences recorded by 15 males and 15 females are used as training set. The total length is 140 seconds. 4 sentences recorded by 2 males and 2 females are used as testing set. The total length is 15 seconds. Five kinds of noise with different frequency characteristics are used. They are machine noise (in high band), siren noise (in medium band), babble noise (in wideband), white noise (in wideband) and factory noise (in low band). The first three kinds of noise with increasing energy level and total length of 140 seconds are used in training and separately added into clean testing speech as matched noise mixture, and the last two kinds are added into clean testing speech as unmatched noise mixture.

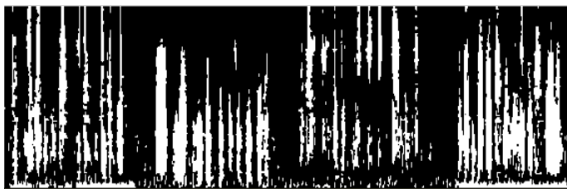

Fig. 4. The binary mask after doing Connected Component Labeling and Hole Filling on Figure.2

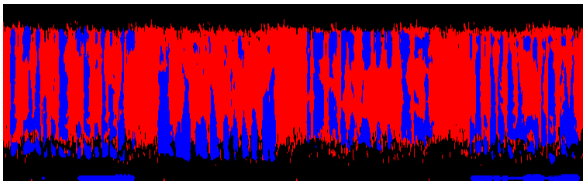

Fig. 5. An example ideal binary mask

4.1 Signal-to-Noise Ratio

MFCCs and pitch are used as features of SVM to determine the binary mask to classify speech units and noise units. The experimental parameters are shown in Table 1.

Different measures are used to evaluate the experimental results. First, on signal level, we use Signal-to-Noise Ratio (SNR) to evaluate. Then, by comparing the ideal binary mask and the mask from our method, several measures are used including HIT-FA Rate (HIT rate minus False Alarm rate) [13,14], which is the difference between Hit Rate (Hit) and False Alarm rates (FA), True Rejection Rate (TRR), True-Acceptance Rate (TAR), Filtering Rate (FR) and Distortion Rate (DR).

Table 1. The Experimental Parameters

Name	Value
sampling frequency	12000Hz
bits/Sample	16bit
frame length	40ms
frame overlapping	20ms
window function	Hamming window
order of Gammatone bank	4
center frequency of Gammatone bank	[50, 6000]
no. of channels of Gammatone bank	128
MFCC feature dimensions	39

Table 2. SNRs in Matched Noise and Unmatched Noise Condition

SNR		-3 dB	0 dB
Matched noise	Babble	7.32 dB	8.15 dB
	Machine	6.52 dB	9.84 dB
	Siren	8.16 dB	10.08 dB
	Average	7.33dB	9.35dB
Unmatched noise	White	5.16 dB	7.52 dB
	Factory	3.58 dB	3.95 dB
	Average	4.37dB	5.73dB

Table 3. True Rejection Rate

Noise		NW	NR	TRR(%)
Matched noise	Babble	2386	93486	97.5%
	Machine	7142	88730	92.6%
	Siren	1320	94552	98.6%
Unmatched noise	White	15708	80164	83.6%
	Factory	47831	48041	50.1%

Table 4. True Acceptance Rate

Gender	SR	SW	TAR(%)
Male	39909	5403	88.1%
Female	42728	2584	94.7%

Table 5. Three Tests to Evaluate the Performance of Connected Component Labeling, Hole Filling, and Morphology

Test	Connected Component Labeling	Hole Filling	Morphology
Test 1	No	No	No
Test 2	Yes	Yes	No
Test 3	Yes	Yes	Yes

Table 6. HIT-FA of 0dB/-3dB Mixture of (a) Test 1, (b) Test 2, (c) Test 3

(a)

Noise		HIT	FA	HIT-FA
Matched noise	Babble	44/22%	19/17%	25/5%
	Machine	68/61%	43/42%	25/19%
	Siren	89/83%	59/57%	30/26%
	Average	67/55.3%	40.3/38.7%	26.7/16.7%
Unmatched noise	White	53/36%	36/29%	18/7%
	Factory	84/81%	40/41%	44/40%
	Average	68.5/58.5%	38/35%	31/23.5%

(b)

Noise		HIT	FA	HIT-FA
Matched noise	Babble	45/23%	20/17%	25/6%
	Machine	69/62%	43/42%	26/20%
	Siren	89/84%	59/57%	31/27%
	Average	67.7/56.3%	40.7/38.7%	27.3/17.7%
Unmatched noise	White	55/37%	36/29%	19/8%
	Factory	85/82%	40/41%	45/41%
	Average	70/59.5%	38/35%	32/24.5%

(c)

Noise		HIT	FA	HIT-FA
Matched noise	Babble	34/14%	14/11%	20/3%
	Machine	59/52%	34/33%	25/19%
	Siren	85/77%	53/51%	32/26%
	Average	59.3/47.7%	33.7/31.7%	25.7/16%
Unmatched noise	White	45/26%	25/18%	20/8%
	Factory	79/75%	30/31%	49/44%
	Average	62/50.5%	27.5/24.5%	34.5/26%

First, we evaluate the classification performance of SVM alone. To do this, the post-processing (Connected Component Labeling, Hole Filling, and Morphology) on binary mask is not added in the experiments of 4.1 and 4.2.

The input sound mixtures with signal to matched noise or unmatched noise ratio of -3dB and 0dB are used in our experiment. After our speech segregation system, speech and noise are separated and the output SNRs are shown in Table 2.

As shown in Table 2, in matched noise condition, the -3dB mixture can improve to the average 7.33dB, and the 0dB mixture can improve to the average 9.35dB. In unmatched noise condition, the -3dB mixture can improve to the

Table 7. FR of 0dB/-3dB Mixture

Noise		Test 1	Test 2	Test 3
Matched noise	Babble	80/82%	80/83%	86/89%
	Machine	56/57%	57/58%	66/67%
	Siren	41/43%	41/43%	46/48%
	Average	59/60.7%	59.3/61.3%	66/68%
Unmatched noise	White	64/71%	64/71%	74/81%
	Factory	59/58%	59/58%	69/68%
	Average	61.5/64.5%	61.5/64.5%	71.5/74.5%

Table 8. DR of 0dB/-3dB Mixture

Noise		Test 1	Test 2	Test 3
Matched noise	Babble	56/78%	55/73%	66/85%
	Machine	33/57%	31/38%	41/59%
	Siren	11/17%	10/16%	14/22%
	Average	33.3/50.7%	32/42.3%	40.3/55.3%
Unmatched noise	White	46/63%	45/63%	54/73%
	Factory	16/19%	15/18%	20/25%
	Average	31/41%	30/40.5%	37/49%

average 4.37dB, and the 0dB mixture can improve to the average 5.73dB.

4.2 True Rejection Rate and True Acceptance Rate

To test the classification performance of SVM, two experiments are set. The input of the first experiment is noise alone and we detect its True Rejection Rate (TRR). The input of the second experiment is clean speech and we detect its True Acceptance Rate (TAR). The TRR is the percentage of noise units a system correctly reject and the TAR is the percentage of speech units a system correctly verifies. In ideal cases, supposedly, we will get 100% TRR and TAR for noise (the first experiment) and clean speech (the second experiment) conditions.

Table 3 present the TRR results of the matched or unmatched noise. NR is the number of correctly

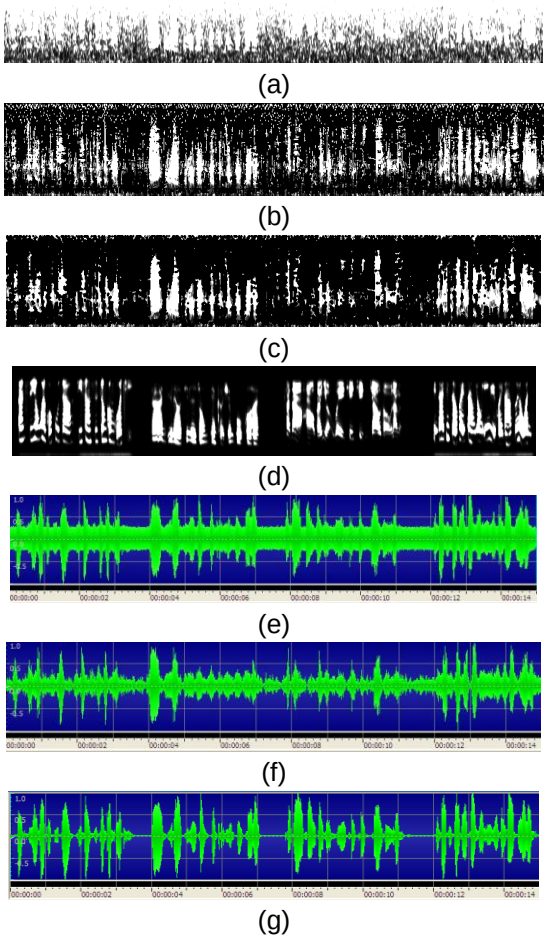

Fig. 7. T-F units of (a) -3 dB mixture with white noise (b) Test 2 (c) Test 3 (d) clean speech. Waveforms of (e) -3 dB mixture with white noise (f) Test 3 (g) clean speech

classified noise units (reject), and NW is the number of misclassified noise units (false alarm). SR is the number of correctly classified speech units (hit), and SW is the number of misclassified speech units (miss). As Shown in Table 3, obviously, it is more difficult to filter out noise correctly when the noise is unmatched (untrained). The result of factory noise, which distributes in low band and is difficult to distinguish with speech, is the worst.

Then, we input clean male (7.5 second long, the two male testing sentences) and clean female speech (7.5 second long, the two female testing sentences) separately. The result of TARs is shown in Table 4. The TAR is higher in female speech.

4.3 HIT-FA, Filtering Rate and Distortion Rate

To further evaluate the performance of Connected Component Labeling, Hole Filling, and Morphology, we design three tests as in Table 5 and compare their results. Test 1 uses our system without Connected Component Labeling, Hole Filling, and Morphology. Test 2 uses Connected Component Labeling and Hole Filling only, and Test 3 uses all of the three.

Several measures are used, including HIT-FA, FR and DR. HIT-FA is the difference between HIT and FA and is useful in predicting the intelligibility of speech synthesized using estimated binary masks [13][14]. The HIT, FA, FR, and DR are defined as:

$$\text{Hit Rate (HIT)}=SR/(SR+SW) \quad (1)$$

$$\text{False Alarm Rate (FA)}=NW/(NR+NW) \quad (2)$$

$$\text{Filtering Rate (FR)}=NR/(NR+NW) \quad (3)$$

$$\text{Distortion Rate (DR)}=SW/(SR+SW) \quad (4)$$

Higher FR and lower DR are desired for speech segregation.

To calculate these measures, we need to compare our estimated binary mask with ideal binary mask. In our experiment, ideal binary mask is defined as:

If both noise and speech energy are very small (< 0.01), the T-F unit will be ignored and not put into calculation.

*Else if speech energy > 0.5 * noise energy, the T-F unit is labeled as speech.*

*Else if speech energy $\leqq$ 0.5 * noise energy, the T-F unit is labeled as noise.*

An example resulting ideal binary mask is as shown in Figure 5. Blue area is labeled as speech and red area is labeled as noise. Black area are units with very small energy and can be ignored.

Table 6 is the result of HIT, FA, and FIT-FA of 0dB/-3dB mixture. The average unmatched noise HIT-FA of Test 3 is the highest, while the average

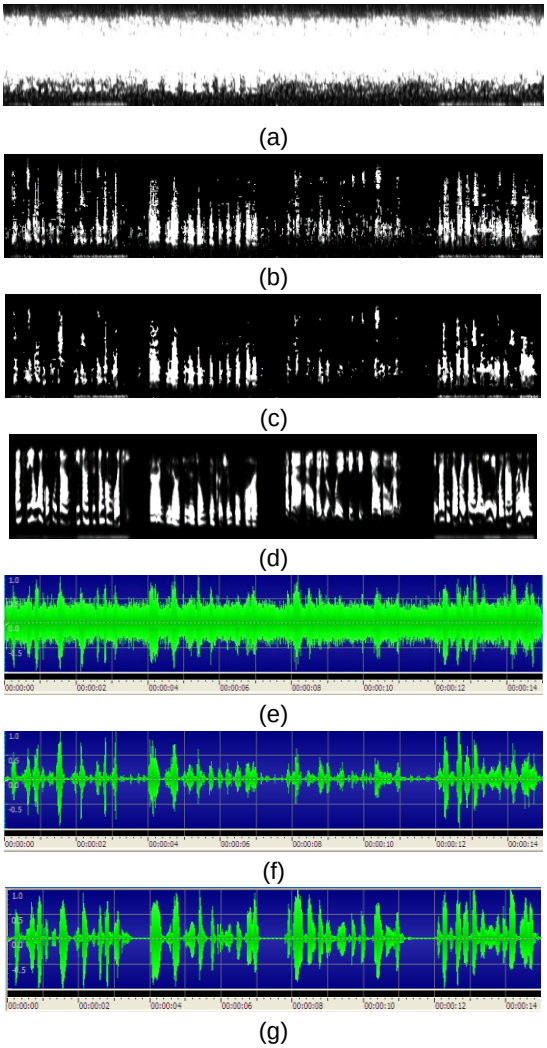

Fig. 6. T-F units of (a) -3 dB mixture with babble noise (b) Test 2 (c) Test 3 (d) clean speech. Waveforms of (e) -3 dB mixture with babble noise (f) Test 3 (g) clean speech

matched noise HIT-FA of Test 2 is the highest. Table 7 and 8 show the FR and DR results of 0dB/-3dB mixture. Comparing the results shown in Table 7 and Table 8, Test 3 has higher FR and Test 2 using Connected Component Labeling and Hole Filling only has lower DR. That is, although Morphology can increase FR, it also increases DR.

The T-F units of -3 dB mixture with matched babble noise and unmatched white noise, the T-F results of Test 2, Test 3, and clean speech are shown in Figures 6 and 7. The waveforms of mixture, Test 3, and clean speech are also shown. Comparing our results with the sound mixtures, our method can successfully segregate speech and improve the speech quality.

5 Conclusions

This paper proposes SVM classification and post-processing including Component Labeling, Hole Filling, and Morphology on CASA mask for speech segregation. By observing the results of different measures, T-F units, and waveforms, our method separates speech from background noise effectively.

Acknowledgements

This work was supported by MOST, Taiwan under Contract MOST 103-2221-E-327 -034.

References

1. **Wang, D.L. & Brown, G.J. (2006).** *Computational auditory scene analysis: Principles, algorithms and applications*. John Wiley & Sons, Inc.
2. **Jang, G.J., Lee, T.W., & Oh, Y.H. (2003).** Single-channel signal separation using time-domain basis functions. *IEEE Signal Processing Letters,* Vol. 10, No. 6, pp. 168–171. DOI:10.1109/LSP.2003.811630.
3. **Yilmaz, O. & Rickard, S. (2004).** Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing,* Vol. 52, No. 7, pp. 1830–1847. DOI:10.1109/TSP.2004.828896.
4. **Ayllon, D., Pita, R.G., Amores, P.J., Zurera, M.R., & Aguilar, C.L. (2011).** Energy-weighted Mean Shift algorithm for speech source separation. *IEEE Statistical Signal Processing Workshop*, pp. 785–788. DOI: 10.1109/SSP.2011.5967822.
5. **Kocsis, Z., Winkler, I., Szalárdya, O., & Bendixend, A. (2014).** Effects of multiple congruent cues on concurrent sound segregation during passive and active listening: An event-related potential (ERP) study. *Biological Psychology*, Vol. 100, pp. 20–33. DOI:10.1016/j.biopsycho.2014.04.005

6. **Wang, D.L. & Divenyi, P. (2005).** On Ideal Binary Mask as the Computational Goal of Auditory Scene Analysis. *Speech Separation by Humans and Machines*, Vol. 12, pp. 181–197.
7. **Brungary, D.S., Chang, P.S., Simpson, B.D., & Wang, D.L. (2006).** Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *Journal of the Acoustical Society of America*, Vol. 120, p.p. 4007–4018.
8. **Han, K. & Wang, D.L. (2011).** An SVM based classification approach to speech separation. *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4632–4635.
9. **He, L., Zhao, X., Chao, Y., & Suzuki, K. (2014).** Configuration-Transition-Based Connected-Component Labeling. *IEEE Transactions on Image Processing,* Vol. 23, No. 2, pp. 943–951.
10. **Mao, Y., Cheung, G., Ortega, A., & Ji, Y. (2013).** Expansion hole filling in depth-image-based rendering using graph-based interpolation. *Acoustics, Speech and Signal Processing (ICASSP),* pp. 1859–1863.
11. **Steel, A. & Brunner, D. (2008).** Detection and Characterization of Urban Objects from VHR Optical Image Data. *Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1256–1259.
12. **Yin, H., Hohmanna, V., & Nadeua, C. (2011).** Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency. *Speech Communication*, Vol. 53, No. 5, pp. 707–715.
13. **Kim, G., Lu, Y., Hu, Y., & Loizou, P.C. (2009).** An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *Journal of the Acoustical Society of America*, Vol. 126, pp. 1486–1494.
14. **Han, K. & Wang, D.L. (2012).** A classification based approach to speech segregation. *Journal of the Acoustical Society of America*, Vol. 132, pp. 3475–3483.

Article received on 21/12/2016; accepted on 22/02/2017. Corresponding author is Wen-Hsing Lai.