Development of Novel Hearing Aids by Using Image Recognition Technology

Bor-Shing Lin, *Member, IEEE*, Ching-Feng Liu, Chih-Jen Cheng, Jhi-Joung Wang, Chengyu Liu, Jianqing Li, and Bor-Shyh Lin, *Senior Member, IEEE*

***Abstract*—Speech is easily affected by different background noise in real environment to reduce the speech intelligibility, in particular, for hearing impaired listeners. In order to improve the above issue, several hearing aids have been developed to enhance the speech signal in noisy environment. Most of current hearing aids were designed to enhance the component of speech and suppress the component of noise. However, it is difficult to separate other speech sources. Adaptive signal enhancement with the beamforming technique might improve the above issue. However, how to distinguish the location of the desired speaker effectively is still a difficult challenge for adaptive beamforming method. A novel concept of hearing aid was proposed in this study. Different from the beamforming-based hearing aids, which use the cross-correlation-coefficient method to estimate time difference of arrival (TDOA), an image recognition technology was used to estimate the location of the desired speaker to obtain the more precise TDOA. An adaptive signal enhancement was also used to enhance the noisy speech sound. From the experimental results, the proposed system could provide a smaller absolute error of TDOA less than 1.25×10^{-4} ms, and a clear speech sound from the target speaker who the user wants to listen to.**

***Index Terms*—Hearing aids, adaptive filter, cross-correlation-coefficient, time difference of arrival, image recognition technology.**

Manuscript received January 29, 2018; revised April 2, 2018 and May 7, 2018. Date of publication May 15, 2018; date of current version May 6, 2019.This work was supported in part by the Ministry of Science and Technology in Taiwan, under Grants MOST 106-2221-E-009-059 and MOST 106-2221-E-305-012, and in part by the University System of Taipei Joint Research Program, under Grants USTP-NTPU-TMU-104-01 and USTP-NTUT-NTPU-106-03. *(Corresponding author: Bor-Shyh Lin.)*

B.-S. Lin is with the Department of Computer Science and Information Engineering, National Taipei University, New Taipei City 23741, Taiwan (e-mail: bslin@mail.ntpu.edu.tw).

C.-F. Liu and J.-J. Wang are with the Department of Medical Research, Chi Mei Medical Center, Tainan 71004, Taiwan (e-mail: wtcen@hotmail.com; 400002@mail.chimei.org.tw).

C.-J. Cheng and B.-S. Lin are with the Institute of Imaging and Biomedical Photonics, National Chiao Tung University, Tainan 71150, Taiwan (e-mail: a79316@gmail.com; borshyhlin@gmail.com).

C. Liu is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: chengyu@seu.edu.cn).

J. Li is with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China, and also with the School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 210029, China (e-mail: ljq@seu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2018.2836180

I. Introduction

IN REAL environment, speech is easily affected by different background noise, and this also reduces the speech intelligibility to cause trouble for the listener, in particular, for hearing impaired listeners [1]. In order to improve the above issue, several hearing aids were developed to amplify the desired speech sounds and reduce environmental noise [2]. Several previous studies attempted to apply single microphone in hearing aids to enhance the interesting speech sound, but their efficiencies of enhancing speech intelligibility seem poor [3], in particular, when the sound spectra of background noise overlap that of the interesting speech sound [4].

In order to improve the above issue, the technique of microphone array was also applied in the design of hearing aids. By using the technique of microphone array, besides temporal and spectral information, spatial information can also be utilized to improve the efficiency of speech enhancement according to the position of speech sources [5]. The delay-and-sum beamforming method is one of the simplest methods used in the technique of microphone array. The delay-and-sum beamforming algorithm estimates the time delays of received signal from each sensor by aligning and summing the input signals of the microphones [6]. Minimum variance distortionless response (MVDR) [7] is one of the most frequently used adaptive beamforming methods, used for adaptively searching the optimum location of the target sound and tracking the variation of the spatial noise field to dramatically improve the performance of the beam former on noise cancellation [8]. The well-known adaptive beamformer of generalized sidelobe canceller (GSC) algorithm was proposed by Griffiths and Jim, and GSC has become one of the basic frameworks of adaptive noise reduction [9], [10]. Here, GSC algorithm consists of a fixed beamformer, a blocking matrix, and a multichannel adaptive noise canceller (ANC) [11]. The fixed beamformer and blocking matrix are used to produce primary sound signals and noise reference signals respectively, and the multichannel ANC is used to eliminate the noise components in the fixed beamformer output. Generalized sidelobe canceller algorithm can provide a better performance on noise cancellation than that of that fixed beamforming techniques, and contains the ability of adapting the acoustic environments [12]. However, the main problem of GSC is that the desired speaker location has to be provided to estimate the time difference of arrival (TDOA) between signals received at spatially separated microphones. How to distinguish the location of the desired

sound source is usually very difficult for the current hearing aids.

Extraction of TDOA between signals received at two spatially separated sensors has been widely applied to sonar and radar to find the target position [13]. Cross-correlation is a standard technique for TDOA estimation in array processing. In order to estimate the TDOA between two microphones, the time domain cross-correlation between two signals has to be calculated, and the lag, that provides the maximum cross-correlation, can be viewed as TDOA. This technique performs well in anechoic environments, but its performance degrades rapidly with increasing reverberation [14].

In order to improve the above issue, a novel hearing aid with an image recognition technology was proposed in this study to provide the more precise TDOA information. Here, a webcam with a wide-angle lens was used to capture the environmental image to detect the location of the desired speaker. The dual microphones were also used to collect noisy speech sounds. Different from other beamforming-based hearing aids, which estimate TDOA by using cross-correlation-coefficient method, the location of the desired speaker could be selected directly and then the TDOA information could be estimated by using the image recognition technology. Finally, adaptive signal enhancement filter (ASEF) was also used to enhance noisy speech sounds. Compared with the cross-correlation-coefficient method, the image recognition technology could provide more precise TDOA information, to ensure that the ASEF could provide a good performance on enhancing noisy speech sounds.

II. System Architecture

In the previous study, the cross-correlation approach is usually used to estimate TDOA. However, in fact, the maximum correlation between noisy speech sounds obtained by a microphone array is not always equal to TDOA, because the location of the interesting speaker is not always nearest to the listener or the speech sound of the interesting speaker is not always loudest, in particular, when many people speak simultaneously. Different from the conventional cross-correlation approach, TDOA is estimated straightforwardly from the location of the desired speaker recognized by a face recognition algorithm. By using the proposed concept of estimating TDOA, it not only can determine the location of the sound source accurately, but also can avoid the interference of reverberation and background noisy sounds. The basic scheme of the proposed hearing aids was illustrated in Fig. 1. It mainly contains a sound acquisition module, a webcam with a wide-angle lens, and a host system. Here, the webcam was placed at the central location of the dual microphones, and was used to capture the environmental image surrounding with the listener. The sound acquisition module with the dual microphones was used to collect noisy speech sounds from the right and left side of the listener, and then transmit the sound data to the host system. In the host system, a real-time monitoring program was designed to recognize the human face from the environmental image, to estimate the location of the desired speaker, and to enhance the desired speech sound from noisy speech sounds. Finally, the enhanced speech sound would be transmitted to the sound acquisition module, and was played by an earphone.

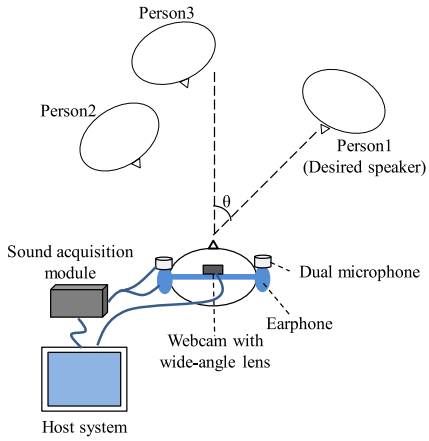

Fig. 1. Basic scheme of proposed hearing aids with image recognition technology.

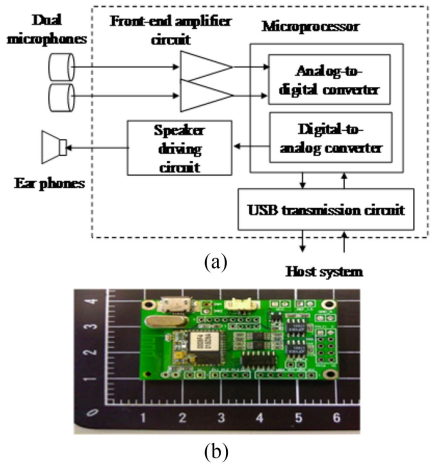

Fig. 2. (a) Block diagram and (b) photograph of sound acquisition module.

A. Sound Acquisition Module

The photograph and block diagram of the proposed sound acquisition module were illustrated in Fig. 2(a) and (b) respectively. The size of the proposed module is about 45 × 30 × 5 mm^3, and it mainly contains several parts, including two microphones, front-end amplifier circuits, a microprocessor, an earphone, a speaker driving circuit, and a USB transmission circuit. First, noisy speech sounds would be collected from the right and left side of the listener by the two microphones. Next, these noisy speech sounds would be amplified and filtered by the front-end amplifier circuits. The front-end amplifier circuits contain pre-amplifiers and band-pass filters. The total gain of the front-end amplifier is set to 40 times with the frequency band of 150 Hz∼1000 Hz. Then, the noisy speech sounds would be digitized by the 12-bit analog-to-digital converter (ADC) built in the microprocessor, with the sampling rate of 20k Hz, and then would be transmitted to the host system via USB interface to perform the processing of adaptive signal enhancement. After the processing of signal enhancement, the enhanced speech sound would be sent from the host system to the sound

acquisition module via USB interface, to generate the analog signals of the enhanced speech by using 12-bit digital-to-analog converter (DAC) with the sampling rate of 20k Hz. Finally, the analog signals of the enhanced speech would be sent to the speaker driving circuit to play the enhanced speech sound via the earphone.

B. Webcam With Wide-Angle Lens

In this study, a webcam with a 120-degree wide-angle lens (WIDECAM F100, Genius, Taiwan), which can provide a better angle of view, was used to capture the environmental image surrounding with the listener. It can provide the image resolution of 12 million pixels, and the frame rate of 30 frames-per-second.

C. Host System

In this study, a commercial tablet with the operation system of Windows 10 was used the platform of the host system. The software architecture of the real-time monitoring program in the host system contains three parts: GUI, BUFFER, and THREAD. The part of graphical user interface (GUI) provides the ability to precisely display and draw of the GUI elements. BUFFER is a link-list container, used to store the received sound data. THREAD contains four independent execution threads, including USB Application Program Interface (USB APT), RECEIVE, TRANSMIT, ASEF and Opencv. Here, USB API was used to connect the host system with the sound acquisition module via USB interface. The thread of RECEIVE was designed to receive sound data obtained from the sound acquisition module, and store them into BUFFER. The thread of TRANSMIT was used to transmit the enhanced speech sound to the sound acquisition module. The thread of ASEF was designed to implement the algorithm of adaptive signal enhancement to enhance noisy speech sounds. Opencv API was used to capture the image and image processing. Fig. 3(a) and (b) show the flowchart and screenshot of the real-time monitoring program respectively. In the beginning of this program, the GUI of this program would be first built to allow the user operating this system. Next, the sound acquisition module would be connected with the host system by the thread of USB API. Then, the thread of Opencv API would be enabled to capture the environmental image via the webcam with a wide-angle lens, and to detect the faces and eyes of the human automatically. By operating the graphical user interface of this program, the user could select the desired speaker from these detected human faces, and then the TDOA of the desired speaker would be estimated. Next, this program would continuously receive the speech data from the signal acquisition module and store into BUFFER. Next, the thread of ASEF would be performed to enhance the received noisy speech sounds. Then, the enhanced speech sound would be transmitted to the sound acquisition module, and would be played via the earphones.

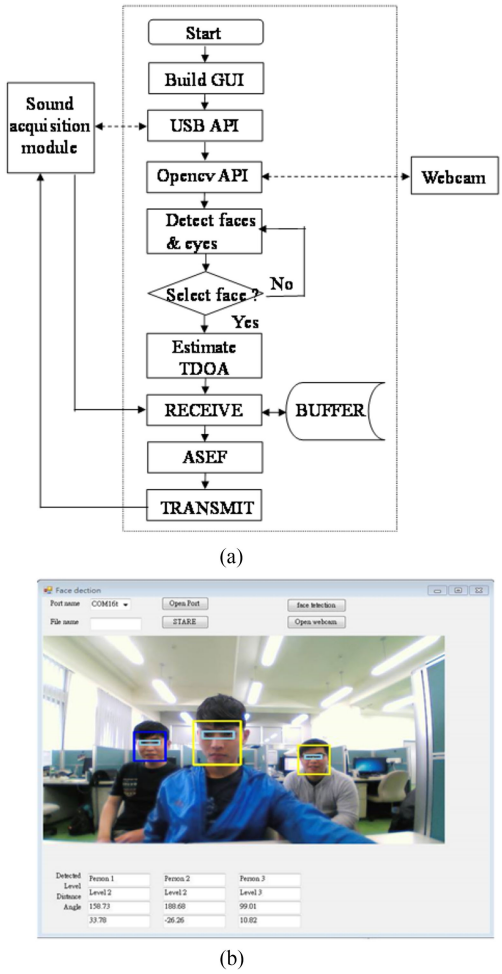

Fig. 3. (a) Flowchart and (b) screenshot of real-time monitoring program.

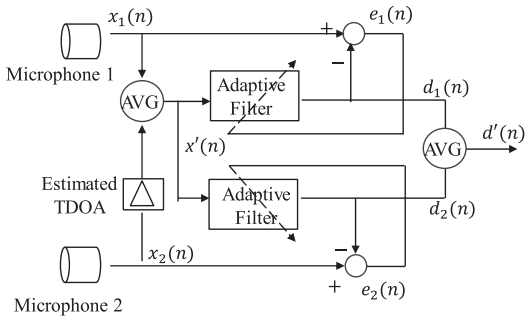

Fig. 4. Basic scheme of adaptive signal enhancement filter in proposed system.

III. Methods

A. Adaptive Signal Enhancement Filter

The basic scheme of adaptive signal enhancement used in the proposed system was illustrated in Fig. 4. Let $X_1(n) = [x_1(n), x_1(n-1), \ldots, x_1(n-M+1)]$ and $\mathbf{X}_2(n) = [x_2(n), x_2(n-1), \ldots, x_2(n-M+1)]$ denote the sequences of noisy speech sounds at iteration n, obtained by different microphones, and they were used as the primary input of the adaptive filters. Next, the beamforming technique would be used to simply improve the signal-to-noise ratio (SNR) of noisy speech sounds. The phase of $\mathbf{X_2}(n)$ would be shifted to the estimated TDOA, and then would be combined with $\mathbf{X_1}(n)$, to result in the constructive combination $\mathbf{X}'(n)$. The signal of $\mathbf{X}'(n)$ would be used as the reference input of the adaptive filters. For the *i*-th M-order

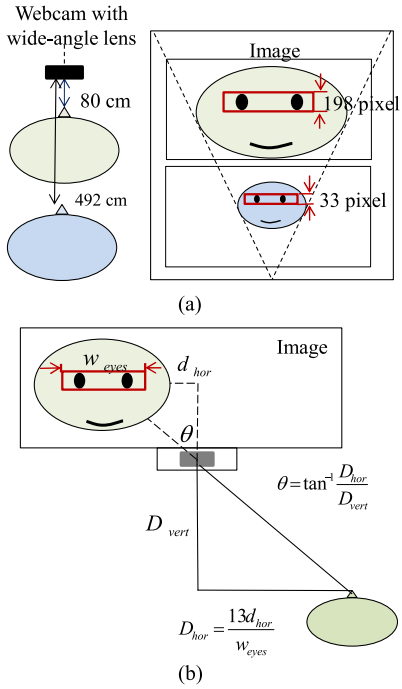

Fig. 5. Illustration for estimating (a) distance and (b) angle of desired speaker.

adaptive filter, its output $d_i(n)$ at iteration n can be expressed as

$$d_i(n) = \mathbf{W_i^T}(n)\mathbf{X}'(n), i = 1, 2 \tag{1}$$

where $\mathbf{W_i}(n) = [w_{i,0}(n), w_{i,1}(n), \ldots, w_{i,M-1}(n)]^T$ denote the $M \times 1$ weight vector of the i-th adaptive filter. In this study, least-mean-square algorithm (LMS) [15], which is a steepest descent method, was used to adapt the filter weight, and they can be given by

$$e_i(n) = x_i(n) - d_i(n) \tag{2}$$

$$\mathbf{W_i}(n+1) = (1-\delta)\mathbf{W_i}(n) + \mu y(n) e_i(n) \tag{3}$$

where δ is the leakage factor limited between 0 and 1, and was set to 0.0001 in this study. Here, μ denotes the learning rate. Finally, the enhanced speech sound $d(n)$ at iteration n can be obtained by calculating the average of $d_1(n)$ and $d_2(n)$.

B. Estimation for Time Difference of Arrival

Open Source Computer Vision Library (OpenCV) is a project, initiated by Intel in 1999, and it provides a number of functions for image processing and computer vision. In this study, the Haar cascade classifier [16] built in OpenCV was used to detect the faces in the environmental image surrounding with the listener. From the size and location of the detected eyes in the face image, the vertical distance and angle of the desired speaker can be estimated to evaluate the value of TDOA. Fig. 5(a) illustrates the estimation for the vertical distance of the desired speaker. In this example, the heights of the eyes are about 198 and 33 pixels when the vertical distances of the desired speaker are about 80 cm and 490 cm respectively. By using the geometry linear equation, the vertical distance D_{vert} of the desired speaker can

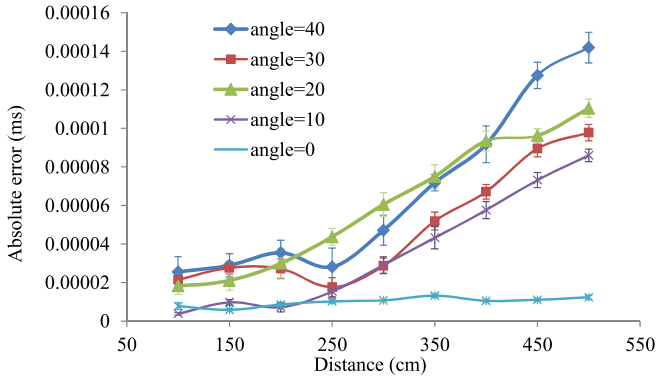

Fig. 6. Performance of estimating TDOA under different distances and angles.

be estimated from the height H_{eyes} of the eyes, and it is given by

$$D_{\text{vert}} = -2.497 H_{\text{eyes}} + 572.745 \tag{4}$$

In order to validate the above equation, the vertical distance of the desired speaker was changed from 70 cm to 500 cm. The mean squared error between the actual vertical distance and the estimated distance of the desired speaker is about 9.16 pixels. Fig. 5(b) illustrates the estimation for the angle of the desired speaker. Before estimating the angle of the desired speaker, his/her horizontal distance has to first be estimated. The averaged width of the actual human eyes is about 13 centimeters. Therefore, if the distance of the desired speaker from the image center is l pixel, then the actually horizontal distance D_{hor} of the desired speaker can be expressed by

$$D_{\text{hor}} = \frac{13 d_{\text{hor}}}{W_{\text{eyes}}} \tag{5}$$

where W_{eyes} is the width of eyes on the image. Then the angle of the desired speaker can be estimated from his/ her horizontal distance

$$\theta = \tan^{-1} \frac{D_{\text{hor}}}{D_{\text{vert}}} \tag{6}$$

Finally, the actual distances between the desired speaker and different microphones can be calculated from the estimated vertical distance and angle of the desired speaker by using Pythagorean theorem. Then, the values of TDOA can be obtained from the actual distance of the desired speaker corresponding to different microphones and the speed of sound under the general environment is 346 m/s as in air at 25 °C.

IV. Results

A. Performance on Estimating Time Difference of Arrival

In this section, the performance of estimating TDOA under different distances and angles by using the image recognition technology was investigated. In this experiment, the distance of the desired speaker from the webcam was set from 1 m to 5 m, and the angle of the desired speaker was set to $0^o \sim 40^o$. Fig. 6 showed the absolute error $|\hat{\tau} - \tau|$ of TDOA corresponding to

different distances and angles. Here, $\hat{\tau}$ and τ denote the TDOA estimation and true TDOA. It indicated that the absolute error would increase with the increase of distance, in particular, when the angle was larger than 0 degree.

B. Performance on Enhancing Speech Sound Under Different Noise Levels

Before evaluating the performances of enhancing noisy speech sounds, the noisy speech sounds with different signal-to-noise ratio (SNR) under different environmental conditions were first generated. The definition of signal-to-noise ratio for the noisy speech sounds is given by

$$SNR = 20\log_{10}\left(\frac{A_s}{A_n}\right) \tag{7}$$

where A_s and A_n denote the root mean square (RMS) of noise-free speech sound and noise respectively. Fig. 7(a) and (b) showed two settings of environmental conditions, the speech sounds of interest (noise-free speech sounds), the noisy speech sounds, and the filtered speech sounds (learning rate = 0.4; ASE filter order = 16). The signal-to-noise ratios of the noisy speech sound in Fig. 7(a) and (b) were about -2.95 dB and -2.2 dB respectively. These two setting conditions were designed after considering. Condition 1 is deliberately on both sides, the angle is relatively large, this setting was used to verify whether it is easier to eliminate noise interference; Condition 2 is intentionally closer, and the noise source is close to the target speech source. The interference of noise source should be larger than condition 1. This setting was used to verify whether noise can still be effectively eliminated, but the effect will be worse than condition 1. From the experimental results, the noise in the speech sounds could be effectively reduced to enhance the speech sound. Next, the performances of ASEF on enhancing speech sound under different noise levels were also investigated. Fig. 8 shows the performance of enhancing speech sounds corresponding to different learning rates. It showed that it could provide a better performance when the learning rate was set to 0.4. Fig. 9 showed the performance comparison of enhancing speech sounds corresponding different noise levels. Here, the approach of mean square error (MSE) between the noise-free speech sound and the filtered speech sound was used to estimate the performance of ASE. The SNR of the noisy speech sound was set from -8 dB to 0 dB. It showed that the value of MSE would increase when the SNR of noisy speech sound became poorer. SNRs in classroom have typically been reported to be in the range, -7 dB to 5 dB [17]. SNRs of other environments such as hospitals, families are usually less than -7 dB. In Fig. 8, under the SNR of -8 dB, the value of MSE is less than 6×10^{-4}. It means our proposed system can perform good performances in different environments such as schools, hospitals, families, etc.

C. Performance on Enhancing Speech Sound Under Non-Stationary Noise

In this section, the performance of ASE on enhancing speech sound under non-stationary Gaussian noise was investigated. In

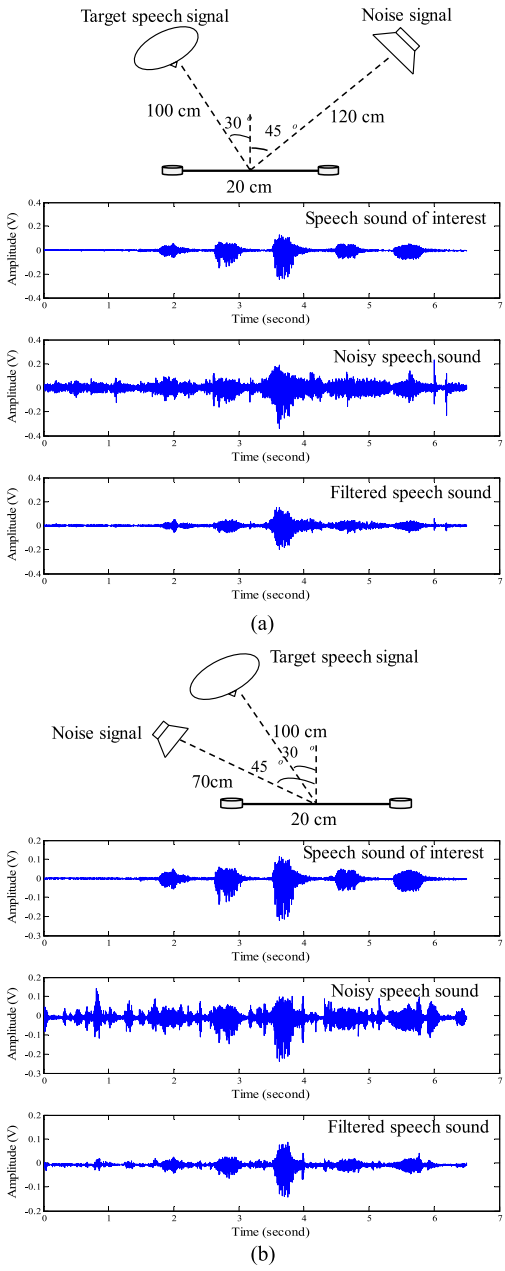

Fig. 7. Speech sounds of interest, noisy speech sounds, and filtered speech sounds under (a) environmental condition 1 and (b) environmental condition 2.

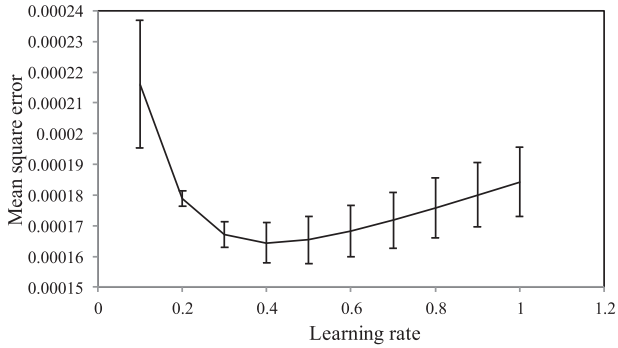

Fig. 8. Performance of enhancing speech sounds corresponding to different learning rates.

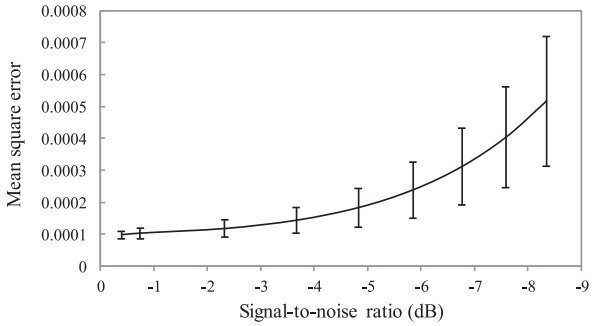

Fig. 9. Performance of enhancing speech sounds under different noise levels.

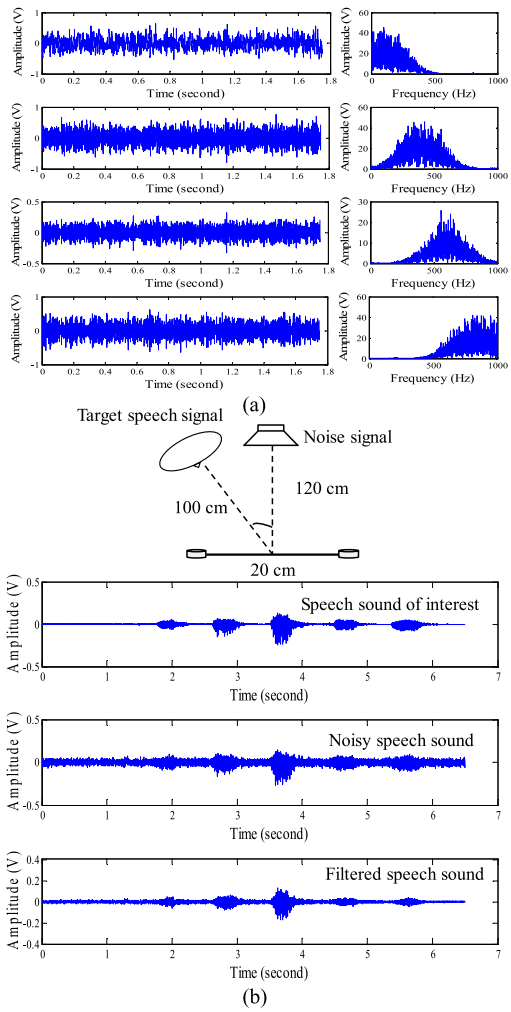

Fig. 10. (a) Four types of Gaussian noises and their power spectrums. (b) Speech sounds of interest, noisy speech sounds, and filtered speech sounds under non-stationary noise.

this experiment, four types of Gaussian noise and their power spectrums, as shown in Fig. 10(a), were used to generate a non-stationary noise. Here, the noise type varied every 1.75 second, and its SNR varied from -5 dB to 0 dB randomly. Fig. 10(b) showed the speech sound of interest, noisy speech sound and filtered speech sound (learning rate $= 0.4$; ASE filter order $= 16$). The experimental result showed that the proposed method could also enhance noisy speech sounds under non-stationary noise.

V. Discussions

Time delay estimation (TDE) is to estimate the relative TDOA from multi-channel sound signals obtained by microphone array. The most popular TDE technique is based on the generalized cross-correlation method, proposed by Knapp and Carter [18]. The estimation of TDE could be viewed as the time-lag that maximized the cross-correlation between sound signals obtained by two microphones. This method performs fairly well in moderately noisy and non-reverberant environments [19], [20]. However, the performance of the generalized cross-correlation method is very sensitive to reverberation and tends to break down when reverberation or noise is high [21]. In order to reduce the influence of noise and reverberation on the performance of estimating TDE, multi-channel cross-correlation-coefficient (MCCC) method was proposed to estimate TDE [22]. The experimental results in this study showed that it caused misjudgment easily with the increase of noise level, when two and three microphones are employed to estimate TDE. In the previous study [23], the TDOA estimation was identified as an anomaly when the absolute error $|\hat{\tau} - \tau| > 1.25 \times 10^{-4}$ ms. It showed that the percentage of anomalies was over 16% under the SNR of -10 dB when two microphones were used. The performance of MCCC method on estimating TDE could be effectively improved by increasing the number of microphones, but this also caused the great increase of computational complexity. In the proposed system, all of absolute errors obtained by using the image recognition technique were less than 1.25×10^{-4} ms when the distance of the desired speaker from the webcam is less than 4 m, and therefore, the percentage of anomalies was approximately zero. The performance of the proposed method on estimating TDOA is independent from the required number of microphones and the SNR of the environmental noise. Therefore, using the image recognition technique to estimate TDOA could effectively improve the issue of the conventional cross-correlation method. The performance of estimating TDOA by using the proposed methods was mainly affected by distance and angle, due to the law of cosines. When the distance of the desired speaker from the webcam became longer and the angle became lager, the value of TDOA also became larger. In this study, ASE algorithm was used to enhance noisy speech sounds. Here, only two microphones were used to collect noisy speech sounds. The average of time-aligned noisy speech sounds was used as the reference signal of ASE to enhance speech sounds. From the experimental results, the performance of ASE on enhancing noisy speech sounds was excellent when TDOA could be effectively estimated.

Several hearing aids have been developed in previous studies, and the system comparison between the proposed system and other hearing aids was listed in Table I. For hearing aid applications, the GSC technique is one of the most popular techniques [24]. It can effectively reduce the computational cost, especially implemented with adaptive algorithms. The GSC algorithm consists of a fixed beamformer, a blocking matrix and a

TABLE I
SYSTEM COMPARISON BETWEEN PROPOSED SYSTEM AND OTHER HEARING AIDS

	Spriet *et al.* [25]	Chan *et al.* [26]	Gil-Pita *et al.* [27]	Proposed system
Method of TDOA estimation	Cross-correlation	-	-	Image recognition technology
Method of speech enhancement	Generalized sidelobe canceller	Decision-direct adaptive gain equalizer	Mel frequency cepstral coefficients, least squares linear classifier	Adaptive signal enhancement filter
Computational complexity	Medium	Medium	High	Low
Functions	Enhancing desired speech sound	Enhancing speech components	Enhancing speech components	Enhancing desired speech sound
Manually selecting desired speaker	No	No	No	Yes
Limitations of use	Cross-correlation is sensitive to environmental reverberation.	It is difficult to separate other undesired speech sources	It is difficult to separate other undesired speech sources	Image recognition technology may be affected by environmental brightness

multichannel ANC. The main function of the fixed beamformer is to align the time delays of the signal received from each sensor and then to sum these time-aligned signals to produce the target sound signals. Then the blocking matrix will obtain the differences between these time-aligned signals, and use them as the reference signal of the multichannel ANC to estimate noise. However, GSC algorithm has to assume the desired speaker location to avoid that the leakage of speech mix in the reference signal to cause the output distortion of the multichannel ANC [25]. In 2014, Chan *et al.* proposed a decision-directed adaptive gain equalizer (AGE) for assistive hearing instruments [26], to boost the speech segments and suppresses noise segments in noisy speech simultaneously. The decision-directed AGE algorithm decomposed the input signal into several sub-bands, and then estimated the SNR of each sub-bands. Next, the decision-directed AGE algorithm would adapt each sub-bands by different gains according to their SNRs. Finally, the enhanced speech sounds could be reconstructed from these gained sub-bands. In 2015, Gil-Pita *et al.* proposed Mel frequency cepstral coefficients (MFCC)-based sound classifiers in digital hearing aids [27]. Here, Mel scale filter bank is a kind of triangular filter, used to extract the characteristics of input signal in frequency domain. Next, least squares linear classifier (LSLC) algorithm was used to classify MFCC into three different listening environments: speech, music, and noise. Then, the speech segment would be enhanced and the noise segment would be suppressed. However, if this method wants to classify more types of listening environments, its computational complexity will increase. Moreover, the above two methods could just enhance some sub-bands and suppress other sub-bands in frequency domains to enhance speech sound, they could not separate the interference contributed from other undesired speech sources. Different from other hearing aids, the proposed system could estimate the more precise TDOA information to provide a good performance on enhancing noisy speech sound. Moreover, the location of the desired speaker could be selected manually from the GUI of the proposed system. At present, there are some limitations on use, the main limitation is that the angle is too large, the effect is relatively poor. The way to solve the problem in the future is to use a higher resolution charge-coupled device with a wider angle a lens; or use a dual lens, one dedicated to a smaller angle, and one dedicated to a wider angle. Another limitation is from image recognition technology. Our proposed system can be used in a multi-person environment, the system can capture the distance and orientation of multiple people at the same time, i.e., the user can specify which speaker to listen to. About the accessories worn on the speaker, wearing hats and glasses will have no effect, because the distance between two eyes is used to calculate the distance from camera. However, if the speakers turn their head, the system does not detect their two eyes, the system will be unable to measure distance from camera to make TDOA failed.

VI. CONCLUSION

A novel concept of hearing aid was proposed in this study. Different from the structure of other hearing aids, by using image recognition technology, the location of the desired speaker could be selected to estimate TDOA effectively. The proposed method could effectively improve the issue of the conventional cross-correlation method, which is sensitive to reverberation of other speech sounds. An adaptive signal enhancement filter was also used to enhance noisy speech sound. By providing relatively precise TDOA, the adaptive signal enhancement filter could provide a stable and good performance on enhancing noisy speech sound. From the experimental results, the proposed system could provide a good performance on enhancing noisy speech sound under different noise levels and non-stationary noise.

REFERENCES

[1] T. C. Smedley and R. L. Schow, "Frustrations with hearing aid use: Candid reports from the elderly," *Hear J.*, vol. 43, pp. 21–27, 1990.
[2] B. Widrow and F. L. Luo, "Microphone arrays for hearing aids: An overview," *Speech Commun.*, vol. 39, no. 1–2, pp. 139–146, 2003.
[3] J. Chen, J. Benesty, Y. A. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.
[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
[5] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1368–1381, Jul. 2011.
[6] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin, Germany: Springer, 2001, pp. 19–38.

[7] B. Widrow *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.
[8] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[9] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Process.*, Berlin, Germany: Springer-Verlag, 2008.
[10] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.
[11] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.
[12] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.
[13] G. C. Carter, *Coherence and Time Delay Estimation: An Applied Tutorial for Research, Development, Test, and Evaluation Engineers*. Piscataway, NJ, USA: IEEE Press, 1993.
[14] K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2156–2164, Nov. 2006.
[15] B. Widrow *et al.*, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.
[16] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, 2004.
[17] C. S. Howard, K. J. Munro, and C. J. Plack, "Listening effort at signal-to-noise ratios that are typical of the school classroom," *Int. J. Audiol.*, vol. 49, no. 12, pp. 928–32, 2010.
[18] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
[19] J. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 6, pp. 998–1003, Dec. 1982.
[20] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar. 1996.
[21] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.
[22] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.
[23] J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Process. Lett.*, vol. 14, no. 3, pp. 157–160, Mar. 2007.
[24] Y. Lee and W. R. Wu, "A robust adaptive generalized sidelobe canceller with decision feedback," *IEEE Trans. Antennas Propag.*, vol. 53, no. 11, pp. 3822–3832, Nov. 2005.
[25] A. Spriet, M. Moonen, and J. Wouters, "Robustness analysis of Multichannel Wiener filtering and generalized sidelobe cancellation for multimicrophone noise reduction in hearing aid applications," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 487–503, Jul. 2005.
[26] K. Y. Chan, S. Y. Low, S. Nordholm, and K. F. C. Yiu, "A decision-directed adaptive gain equalizer for assistive hearing instruments," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 8, pp. 1886–1895, Aug. 2014.
[27] R. Gil-Pita, D. Ayllón, J. Ranilla, C. Llerena-Aguilar, and I. Díaz, "A computationally efficient sound environment classifier for hearing AIDS," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2358–2368, Oct. 2015.