34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Multi-microphone Adaptive Array Augmented with Visual Cueing

Paul L. Gibson, *Member, IEEE*, Dan S. Hedin, *Member, IEEE,* Evelyn E. Davies-Venn, Peggy Nelson and Kevin Kramer

Abstract— **We present the development of an audiovisual array that enables hearing aid users to converse with multiple speakers in reverberant environments with significant speech babble noise where their hearing aids do not function well. The system concept consists of a smartphone, a smartphone accessory, and a smartphone software application. The smartphone accessory concept is a multi-microphone audiovisual array in a form factor that allows attachment to the back of the smartphone. The accessory will also contain a lower power radio by which it can transmit audio signals to compatible hearing aids. The smartphone software application concept will use the smartphone's built in camera to acquire images and perform real-time face detection using the built-in face detection support of the smartphone. The audiovisual beamforming algorithm uses the location of talking targets to improve the signal to noise ratio and consequently improve the user's speech intelligibility. Since the proposed array system leverages a handheld consumer electronic device, it will be portable and low cost. A PC based experimental system was developed to demonstrate the feasibility of an audiovisual multi-microphone array and these results are presented.**

I. Introduction

According to the most recent MarkeTrak survey of 80,000 households, Better Hearing Institute (BHI) estimates that 31.5 million Americans have hearing loss. Hearing loss affects 1 in 10 Americans and 1 in 4 households [1]. Furthermore while 95% of individuals with hearing loss could be successfully treated with hearing aids, only 23% currently use them [2].

The problem of listening to multiple speakers in environments with competing speech babble noise is well known to the hearing aid industry, and, in fact, is not only a problem to hearing impaired (HI) but also to normal hearing (NH) listeners. Common listening situations with competing speech babble noise include restaurant conversations, conference table meetings, and automobile conversations. The primary complaint of HA users is their difficulty understanding speech in noise. Only 51% of hearing aid users report they are satisfied with their hearing aids in noise [3]. Listeners with sensorineural hearing loss (SNHL) show variable performance on tests of speech recognition in noise [4] with some HI listeners performing like normal hearing listeners, and others demonstrating much poorer performance. Both elevated thresholds and supra-threshold distortions [5] have been proposed as contributing to this variability in performance, but thorough explanations have been lacking. Even the most modern hearing aids do not entirely solve this problem, which contributes to complaints by wearers and ultimately rejection of hearing aids.

It has been shown that directional microphones can improve speech intelligibility in noisy environments as long as the desired speaker is not located near the noise source and the listening environment is not highly reverberant. However, these assumptions do not usually apply to many real life situations. Recent reviews by Kompis and Diller [6] and Chung [7] indicated that first-order directional microphones provide about 3 to 5 dB improvement in typical rooms, with smaller improvements in reverberation. Furthermore, two-microphone directional systems do not have sharp enough directional focus to solve the speech babble noise problem. They also require the wearer to be facing the desired sound source. The newest digital noise reduction algorithms are mainly suitable for eliminating non-speech-like babble background noise -but there still remains a need for improvements that adequately remove background speech babble noise.

High-order adaptive microphone arrays that can be steered to the appropriate speaker can solve the speech babble noise problem by rejecting sounds not coming from the intended speaker. Hence, the problem shifts to identifying the speaker in order to accurately steer the array. Our prototype system solves this problem by utilizing a camera with face recognition to steer the array to the desired speaker.

II. Prototype Design

A hardware prototype was developed to demonstrate the system components. The prototype was designed to interface to a laptop computer via USB and was controlled with a Windows software application. The application contained the elements that will be implemented on a smartphone. The prototype contained DSP hardware to perform the microphone array processing along with hardware to interface the array with the PC via USB. The prototype's DSP is the version that will be integrated into the accessory while the rest of the components are PC based for engineering efficiency and will be eventually replace by a smartphone. Fig. 1 shows a block diagram of the hardware components.

This work was supported in part by the U.S. National Institutes of Health Grant 1R43DC011468-01.

K. Kramer is with Advanced Medical Electronics, Maple Grove, MN 55369 USA (phone: 763-515-5315; fax: 763-463-4817; e-mail: kkramer@ame-corp.com).

D. Hedin is with Advanced Medical Electronics, Maple Grove, MN 55369 USA (phone: 763-515-5335; fax: 763-463-4817; e-mail: dhedin@ame-corp.com).

P. Gibson is with Advanced Medical Electronics, Maple Grove, MN 55369 USA (phone: 763-515-5360; fax: 763-463-4817; e-mail: pgibson@ame-corp.com).

P. Nelson is with Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN 55455 USA (phone: 612-625-4569; fax: 612-624-7586; e-mail: nelso477@umn.edu).

E. Davies-Venn is with Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, MN 55455 USA (phone: 612-624-0549; fax: 612-624-7586; e-mail: davi0619@umn.edu).

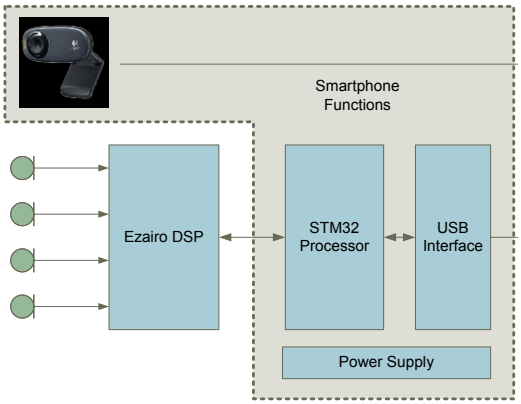

Figure 1. Prototype hardware block diagram.

The PC application captures images from the prototype's USB web camera and performs face detection. Once a face was chosen, the software computes the steering angle and communicates this angle to the array processing DSP via a USB port.

A. Hardware

The prototype system uses the Ezairo 5900 Series DSP by ON Semiconductor. It is a state-of-the-art design for high performance and low power hearing aid applications. It has an ultra-low power consumption of less than 1mA at full processing power. The Ezairo processor has 4 audio input channels each with their own programmable amplifier, anti-aliasing filter, and 18 bit analog to digital converter (A/D).

The prototype has a STM32 ARM processor, USB interface and power supply to provide a PC interface. The STM32 Processor is a 32-bit ARM cortex-M3 processor by ST Microelectronics. The STM32 interfaces to the DSP via a 3-wire serial bus. The DSP streams real-time processed audio to the STM32 processor which then sends the data to the PC via the USB. The USB interface was realized with a FT245R USB FIFO interface by FTDI Chip. The prototype device contained a power supply providing the proper voltage levels required with the device being powered from the USB port. A command interface between the PC application and prototype hardware was also implemented to allow control of the settings.

The microphone array was prototyped with the same microphones and arrangement as the final design. It is a linear 4-microphone broadside array. The microphones chosen were low-cost miniature omnidirectional electrets condenser microphones by PUI Audio, Inc. The spacing of the microphones is 33 mm apart making for an array length of 99 mm. The array spacing was determined based on current smartphone form factors. For example, the current iPhone 4S has a height of 115.2 mm.

B. Beamforming Algorithm

The prototype hardware used the well-established technique of delay-sum beamforming to steer the array to the desired angle selected by the user via the face detection. The delay-sum beamforming is ideally suited for the Ezairo DSP since the front-end built-in 18-bit A/D converters has built-in user controllable sample delay registers. The front-end 4-channel analog block (anti-aliasing, pre-amp, and A/D converter with sample delay) was used to sample the microphones directly. The input controller automatically moves that data to an input FIFO for processing. The microphones were sampled at a 16 kHz rate. The front-end preamplifiers contains user programmable hardware gain of between 12-30dB in steps of 3dB. The PC application allowed for adjustment of gain.

The software application contained a module to calculate the required delays between microphones. The delay for a given steering angle equals $ndcos(\theta')/c$ where θ' is steering angle. The first microphone in the array will have a delay of 0, the second will have a 1 delay unit, the third will have 2 delay units, and so on. Once a delay unit was calculated for the given steering angle, it was implemented for each microphone by 3 delay controls. They are full samples, and a course and fine control which are delays implemented in the A/D converter that are set by registers. The respective delay step sizes are 62.5 µs, 7.8125 µs, and 390.625 ns.

The Ezairo firmware receives the calculated delay elements from the PC via the command interface. Each of the A/D course and fine step delay registers would be programmed for each channel. Each channel had a history buffer to allow for sample delay. Pointers to this history buffer are adjusted based on the sample delay parameter. Once the system was configured for a given steer angle the data sample from each microphone channel was summed and then scaled by 4 to maintain unity gain. This audio stream was then streamed to the PC for playback and analysis.

C. Software

The prototype software consists of a calibrator, a tester, a face detector, and a beam steerer. Advanced Medical Electronics (AME) wrote the software in the C# computer language using Microsoft's Visual Studio 2010.

The face detector processes the camera's video stream to determine the location of faces. The system does not need to perform the more difficult task of recognizing the person to whom the face belongs, i.e., face detection not face recognition. AME used a third party Haar classifier to detect faces in video frames. Using a calibration table, the face detector transforms face locations to steering angles. Given a steering angle, the beam steerer computes the microphone delays and downloads them to the array. The calibrator correlates the location of a face with a sweep angle for a given source frequency. The tester allows the user to (1) set the hardware gain, (2) use the prototype device as a single microphone or as a microphone beamforming array, and (3), steer the beam manually or via face detection.

D. Calibration Procedure

Calibration is a one-time procedure that needs to be completed as part of the system design. The production prototype will be calibrated in a controlled environment. This calibration will be valid for all subsequent devices and the system will require no field calibration by users. To calibrate the prototype face detector and microphone array, AME constructed a "target" by mounting a picture of face above a speaker. The speaker was connected to a signal generator in order to create tones between 500 Hz and 5000 Hz. In order to make the face location independent of the camera's resolution, the face detector normalizes the face

location by setting the video frame's left and right edges to zero and unity respectively. We chose the midpoint between the person eyes as the face's location. To conduct a calibration test, the following procedure was performed. First, the target was positioned at a known distance and angle from the array. These quantities were measured by attaching a string to the center of the array and fastening it to the center of the target. The distance was measured with a tape measure and the angle with a compass. Second, the signal generator was set to a specified frequency. Third, the calibrator module steered the beam from 0° to 180° and measured the Root-Mean-Squared (RMS) value of the array's response. Lastly, the sweep angle that generated the largest response was determined, which we refer to as the Maximum Sweep Angle (MSA). AME conducted over one hundred calibration tests in an uncontrolled environment. The distance was varied between three and ten feet in one foot increments. The angle was varied between 30° and 150° in 15° increments. The source frequency ranged between 500 Hz and 5000 Hz.

The calibration results were analyzed to determine the relationship between the target's measured angle and the computed target location. As expected, the relationship is linear and the coefficient of determination is 0.999. Target locations of zero and unity correspond to target angles of 40° and 140°respectively. Hence, the prototype's Field of View (FOV) is 100°. The calibration results were also analyzed to determine the relationship between the maximum sweep angle and target location. This calibration is required because the camera's axis is translated by a couple of inches from the array's axis. The prototype was purposely built in this manner because smartphones typically position their cameras towards a corner, not in the center of the device.

III. Testing Methods

A. Recruitment

Two listeners with normal hearing and two listeners with mild-moderate sensorineural hearing loss were tested using the prototype system at the University of Minnesota under IRB approval. Listeners ranged in age from 22 to 56 years (M = 39, SD = 18.5). Listeners were recruited from an existing pool of research participants at the University of Minnesota. All participants were compensated on an hourly basis for their time. The two listeners with hearing loss had pure tone average (PTA) thresholds between 27 to 30 dB HL. Sensorineural hearing loss was defined as the absence of an air-bone gap greater than 10 dB HL from 500 Hz to 4000 Hz, and normal tympanograms [8]. The listeners with normal hearing loss had PTA thresholds from 2 to 3 dB HL. Standard audiologic testing was administered at the Julia M Davis Speech and hearing clinic at the University of Minnesota. All testing was conducted in a sound isolating booth using Sennheiser headphones.

B. Objective Assessment Methods

Objective testing was conducted at 0 degrees azimuth in direct view of the test signal. Noise stimuli were presented at 90 and 270 degrees azimuth. The phase I prototype allowed for audio capture from 1 microphone (1mic) for comparison to the 4-microphone audiovisual array (4mic). The phase I prototype was set to 0 degrees azimuth and thresholds were determined for 1 microphone (no beamforming) and 4 microphone audiovisual array. Speech reception threshold in noise was measured using the hearing in noise test by Nilsson et al.[9], and Institute of Electrical and Electronic Engineers (IEEE) sentences. The speakers and prototype device were positioned at ear level and an equidistance of 1 meter. The pre-amplifier gain was get to 15dB for all test subjects.

Subjects' signal to noise ratio threshold was measured using an adaptive test procedure. SNR thresholds were determined using the HINT sentences. This task consists of 250 words digitally contained in 50 phonetically balanced lists of ten sentences spoken by a male talker. The lists are equated for length, naturalness and intelligibility. Participants were instructed to repeat everything they heard, even if they only heard part of a sentence. Two sequential 10-sentence lists were presented. The listeners were given 4 practice sentences to become familiar with the task. Four dB steps were used between sentences 1-4 and 2 dB steps were used for sentence 5-20. Starting with sentence 5, if the participant's response was correct, the presentation level was decreased by 2 dB. If the listener's response was incorrect, the presentation level of the sentence was increased by 2 dB. The noise was presented at a fixed level of 65 dB HL in sound field. The threshold was calculated by averaging the presentation levels for the sentences in two 10-sentence HINT lists. The calculated threshold followed the recommended HINT scoring procedure. The individual test scores were compared to the HINT's normative rankings for speech in noise at a zero degree azimuth. The IEEE sentences represent conversational speech with correct syntax, but they are void of any useable context cues. The sentences are organized in 72 lists of ten sentences. Speech reception threshold in noise were measured using a similar protocol as that used for the HINT sentences, using IEEE sentences spoken by a female talker.

C. Subjective Assessment Methods

Subjective assessment was measured using speech by a female and male talker. Listeners were asked to listen to 1 minute segments of an audio book reading and then rate the signal for intelligibility, pleasantness, ease of listening and effort. Listeners were tested comparing the one microphone to the four microphone audiovisual array. The face detection software was assessed subjectively by having listeners face away from the speaker with the test signal and listen through the headphones. The listener held the prototype device for three test scenarios – straight, rotated to the right, and rotated to the left. For each test the listener would rotate and the test source face would be selected via the face detection selecting an angle. The PC application would then calculate the steering angle and download the parameters to the prototype device in the cases when the 4 microphone audiovisual array was being utilized. The speech was presented at 0 degrees and the noise was presented at 90 and 270 degrees azimuth. For each category the listener gave a rating in the range of 1-5 with 5 being the best.

IV. Results

A. Objective Assessment

The results of this test are presented in Fig. 2. These results show that for both the IEEE sentences case and the HINT sentences case, the 4-microphone beamforming

configuration produced lower SNR thresholds than the 1 microphone configurations. In other words the user could understand the speech at lower volumes given a fixed amount of noise. This held for both the normal and hearing loss listener's with the normal hearing individuals performing better than listeners with hearing loss.

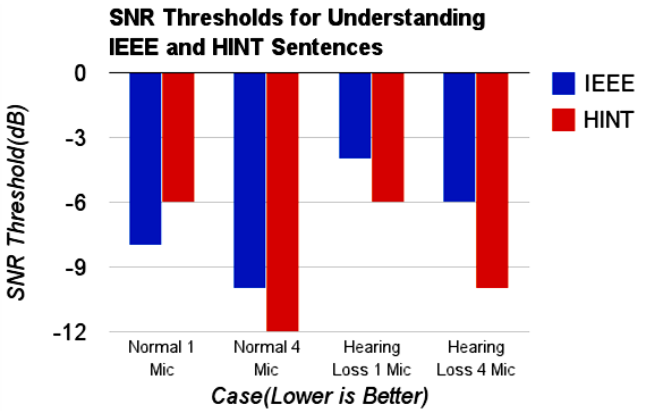

Figure 2. Mean SNR thresholds for IEEE sentences comparing using a single microphone versus the 4 microphone beamforming array.

B. *Subjective Assessment Results*

Fig. 3 and Fig. 4 summarize the comparison results of the subjective testing at 0dB SNR. In Fig. 3, the results for the fixed or non-steering array are shown. In the normal hearing case there is almost no improvement with the array. In the hearing loss case there is some improvement.

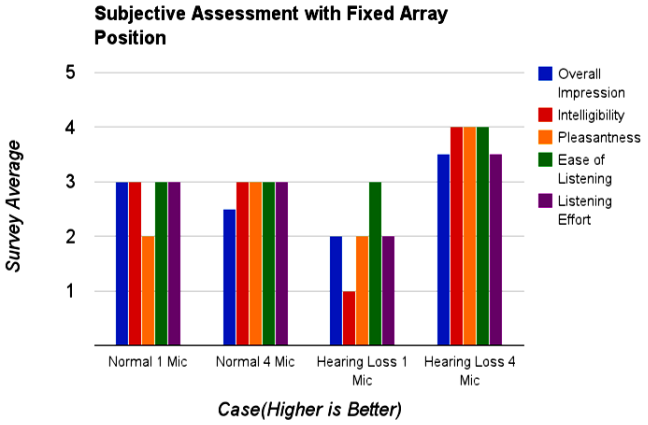

Figure 3. Comparison ratings of using the 4 microphone array configuration verses a single microphone for 0dB SNR. The 4 microphone audiovisual array was always aimed at 90° (straight ahead).

In Fig. 4, the results of the assessment are shown for the case where the array is steered by face detection. The results are split into groups of 1 microphone, 4 microphones, and with and without hearing loss. The results are similar for the single microphones cases for both the normal and hearing loss individuals. The 4 microphone steered array shows much better results and the hearing loss individuals showed more gain than the normal hearing individuals. These results also show that improvement is much greater in the case where the array is steered via face detection than when the array is in a fixed position.

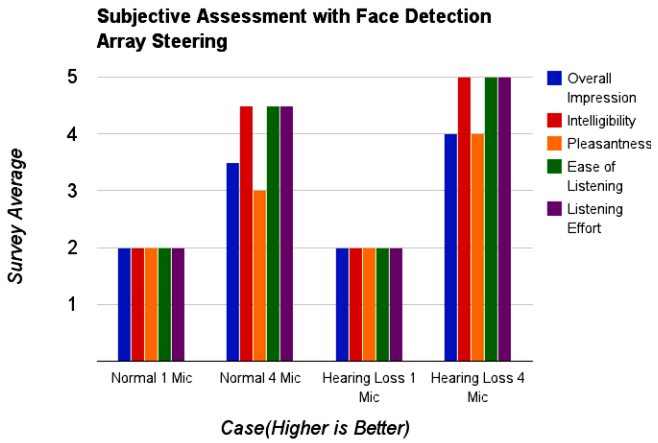

Figure 4. Comparison ratings of using the 4 microphone array configuration verses a single microphone for 0dB SNR. The face detection capability of the system was utilized to steer the 4 microphone audiovisual array.

V. SUMMARY

The testing conducted demonstrated the feasibility of a smartphone-based 4 microphone audiovisual array with visual cueing. The prototype device was shown to be favored over a single microphone in all the test scenarios. Furthermore face detection was shown to be an effective way to steer the microphone array. Future work will include integrating the processing and control into a smartphone application and developing the accessory case.

ACKNOWLEDGMENT

Advanced Medical Electronics thanks the National Institutes of Health for the funding made available to develop and evaluate this novel research tool.

REFERENCES

[1] S. Kochkin, "Prevalence of Hearing Loss.", http://www.betterhearing.org/hearing_loss/prevalence_of_hearing_loss/index.cfm, Accessed 4/2/10.

[2] S. Kochkin, "MarkeTrak VII: Customer satisfaction with hearing instruments in the digital age," *The Hearing Journal*, vol. 58, no. 9, p. 30, 2005.

[3] S. Kochkin, "MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids," *The Hearing Journal*, vol. 60, no. 4, p. 24, 2007.

[4] M. C. Killion and P. A. Niquette, "What can the pure-tone audiogram tell us about a patient's SNR loss," *Hear J*, vol. 53, no. 3, pp. 46–53, 2000.

[5] R. Plomp, "Auditory handicap of hearing impairment and the limited benefit of hearing aids," *The Journal of the Acoustical Society of America*, vol. 63, p. 533, 1978.

[6] M. Kompis and N. Dillier, "Performance of an adaptive beamforming noise reduction scheme for hearing aid applications. II. Experimental verification of the predictions," *The Journal of the Acoustical Society of America*, vol. 109, p. 1134, 2001.

[7] K. Chung, "Challenges and recent developments in hearing aids. Part I. Speech understanding in noise, microphone technologies and noise reduction algorithms," *Trends Amplif*, vol. 8, no. 3, pp. 83–124, 2004.

[8] J. Jerger, "Clinical experience with impedance audiometry," *Archives of Otolaryngology- Head and Neck Surgery*, vol. 92, no. 4, p. 311, 1970.

[9] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *J Acoust Soc Am*, vol. 95, no. 2, pp. 1085–1099, 1994.