A Direction Finding – Beam Forming Conference Microphone System

Michael Steele
Sr. Research Engineer

Lexington Center/School for the Deaf
30th Avenue and 75th Street
Jackson Heights, NY 11370

This paper describes a conference microphone system designed around a TMS320C44, floating point, digital signal processor (DSP). The conference microphone is a real time audio processor designed to perform two primary functions; acoustic beamforming and automatic sound source localization. Although the system was designed specifically as a front end for assistive listening systems, it has many practical applications requiring high quality, speech band audio such as remote transcription services and video-conferencing.

The Direction Finding, Beam Forming (DFBF) conference microphone consists of a tabletop unit containing a 16 microphone array and amplification electronics. This unit is connected to a PCI plug-in DSP card and 16 channel analog to digital (A/D) converter via a 37 pin 'D' type cable. The tabletop unit has a single line level output which can be connected to a variety of transmission and/or amplification systems. While the current implementation uses a personal computer (PC) and a floating point DSP, it is expected that a commercially viable solution would contain a fixed point DSP board and would be completely housed in the tabletop unit.

The acoustic beamformer is a two-band superdirective system using fixed coefficients and 8 microphones. There are two identical arrays arranged at right angles to each other. The geometry of the arrays allows the system to generate four directional patterns, each covering 90 degrees in the horizontal plane. Sound source localization is performed using a modified cross-correlation algorithm which uses periodic features of voiced speech to obtain accurate estimates for talker location. Direction estimates are used to steer the beamforming arrays. The localization algorithm depends on the periodicity of voiced speech and is inherently resistant to non-speech sounds such as door slams or paper shuffling.

1. Background

Conference microphones are designed specifically to receive acoustic speech signals from multiple sources and locations, and to combine these signals into a single output. Two major factors affecting the quality of the output signal are ambient noise and reverberation. Output signal quality is also affected by the method used to receive and combine the acoustic signals. While neither the background noise nor reverberation can be controlled by the system directly, their effects can be minimized by carefully selecting the method used to receive and process the acoustic signals. Common methods include connecting a single microphone to an automatic gain control (AGC), switching between multiple directional microphones, or using an array of microphones combined with adaptive signal processing. These methods vary dramatically in terms of their effectiveness (performance), and computational requirements (cost).

Applications for conference microphones include tele-conferencing, remote transcription services, and assistive listening systems. While all of these would benefit from a microphone which entirely eliminates the effects of background noise and reverberation, no system can eliminate these effects under all possible conditions. Additionally, certain applications will tolerate different levels of signal degradation before becoming unusable. For example, consider a tele-conference, where interactive participants are present at separate locations. Some loss of audio information may be tolerated since attendees can request that a particular sentence, or passage, be repeated. This type of two-way transmission system is known as full-duplex. If two-way video were added to the tele-conference, participants would also receive visual cues. These visual cues would provide additional contextual information, further reducing the quality requirement for the audio channel. Although severely degraded audio is undesireable, it is likely that little information would be lost over the course of a meeting.

One-way transmission, or half-duplex, applications such as remote transcription are more difficult than the full-duplex arrangement. Transcription services generally do not enjoy the luxury of repeat transmission requests described above, or the additional information provided by a

video channel. A remote transcription system typically involves an individual at a remote site who receives audio from the conference microphone (via telephone) and transcribes the received information into text. If the audio received by the transcriber is significantly degraded, there is virtually no method for retrieving the lost information. The problem is further complicated if the transcriber is a computer driven speech recognition system instead of a live person. Studies have been performed at Lexington Center's Rehabilitation Engineering Research Center (RERC) on Hearing Enhancement to evaluate the use of conference microphones as the primary audio source for computer aided remote transcription (CART) services. These studies include comparing transcription errors for local versus remote transcribers, and comparision of speech quality for various commercial conference microphones using the Speech Transmission Index (STI)[5]. Results from these studies can be found in the grant application for this project [1].

A common complaint of individuals who use hearing aids, or other assistive systems, is that they cannot use acoustic amplification devices in noisy or reverberant environments. This is particularly disturbing for people who rely on assisitive listening systems as their primary method of communication. It has been shown by Plomp [8] that hearing impaired individuals require significantly higher speech to noise ratios than hearing listeners. More than 25 million people in the United States have some kind of hearing loss [10]. Of these, slightly less than 4 million use a hearing aid for oral communication, an estimated 175,000 use text telephones (TTY's), and approximately 564,000 use 'other hearing technology' [7]. Due to the widespread use of assistive technologies, there is an obvious necessity to provide high quality speech for transmission systems such as FM, infra-red (IR), and inductive loops (T-coil), which are used widely in large area meetings. Further, as speech recognition technology becomes more frequently used as a front end for assistive devices, it will become increasingly important to provide even higher quality audio since these systems are particularly susceptible to the adverse effects of noise and reverberation.

The DFBF project's primary goal was to develop a portable, DSP based, conference microphone which would be evaluated in real environments and in actual meeting situations. Using specially designed microphone arrays, coupled with an adaptive steering component, it is expected that the DFBF microphone will provide significantly higher speech to noise ratios as compared to currently available commercial microphone systems. At the time of this writing, Lexington Center/School for the Deaf has built two prototype DFBF systems. The prototypes are currently being evaluated. Final results from this study will be available early in 2001, although preliminary results look very promising.

2. System Overview

The DFBF conference microphone hardware consists of two major components; a tabletop microphone array which converts acoustic signals into electrical signals, and a plug-in DSP board responsible for filtering and combining these signals into a single output.

i. The tabletop unit is a finished hardwood enclosure containing 16 electret microphones, two 8 channel preamplifier circuit boards, and an adapter board used to interconnect the various components. The microphones are Primo EM-127 omni-directional electrets. The two preamp boards, designed at Lexington, contain 8 identical amplifier/filter stages used to condition the microphone signals. The circuit for each microphone consists of a high gain (Av = +33dB) preamp followed by a single second-order low-pass filter (Butterworth, Av = +3.9 dB, Fc = 7KHz). Power for the tabletop unit is supplied by an external DC adapter providing +/-5VDC through a 5 pin DIN connector. The output level from each amplifier channel is approximately 1VRMS at a sound pressure of 74dBSPL. The adapter board connects the 16 channels to a single 37 pin 'D' type connector which is used to send the microphone signals to the DSP board. The 37 pin cable contains an additional connnection for returning the DFBF output back to the tabletop unit. This output is accessible through a ¼" audio jack. An additional ¼" audio output is connected to one of the centrally located microphones for monitoring purposes.

ii. The DSP board is a PCI plug-in card purchased from Innovative Integration of Westlake Village, CA. The board contains a TMS320C44 DSP chip running at 60MHz, 128 K local SRAM, 512 K of global SRAM, various peripherals, and two Omnibus mezannine sites used for input/output (I/O) not provided by the DSP board. One of the Omnibus sites contains an SD16 module which is the A/D and D/A interface to the tabletop unit. The SD16 contains 16 sigma-delta codecs capable of providing 18 bit, 96KHz sampling. The other Omnibus site contains a custom board, designed at Lexington, which provides additional digital I/O for testing and controlling the array.

Software components for the DFBF system include the acoustic beamformer, sound source localization, or direction finding, component, and a scan based rotation control which serves as the interface between the beamformer (BF) and the direction finder (DF). Prior to integrating the hardware, a significant amount of analysis and simulation was performed using Simulink (MatLab - Dynamic System Modelling), and programs written in C/C++. After extensive analyses of these software components, the tabletop unit was built and integrated with the DSP. Software was then converted from the simulation models and C/C++ code segments into source code consisting of a mixture of ANSI-C and TMS320C44 assembly language.

3. Beam Former Subsystem

In the presence of diffuse noise fields or room reverberation, which is in effect multi-path propagation of sound, the ability of directional microphones to attenuate signals arriving from other than the desired direction offers a distinct advantage over omni-directional microphones. This advantage is exploited by many products such as hearing aids and assistive listening devices used primarily for receiving and amplifying speech.

Microphone arrays often have different physical geometries although they are usually described by a model similar to the one shown in Figure 1. The array output (1) is a weighted sum of the microphone signals, where the complex weights, a_n, are designed to delay the microphone signals by a prescribed amount. The microphone signals, x_n, can be modeled as delayed replicas of the source where the amount of delay depends on the relative physical location for a given microphone. The equation in (2) describes these signals for a linear, uniformly spaced, endfire array. The phase, or acoustic delay, for each microphone, is a function of the source frequency, ω, the propagation angle of the source, Θ, and the relative position of the microphone. The beamformer system uses the applied weights to emphasize a particular spatial direction and maximize the output signal to noise ratio (SNR).

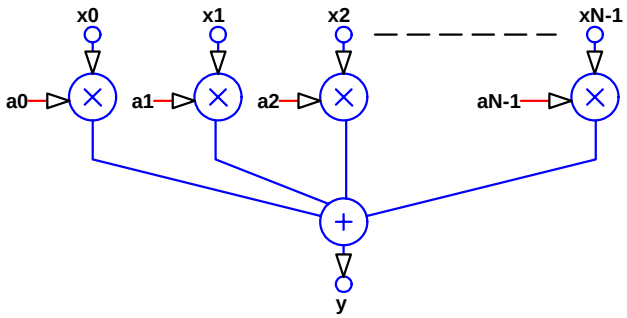


Figure 1. Basic Diagram for an Acoustic Beamformer

$$y(\omega,\Theta)=\sum_{n=0}^{N-1}a_n x_n \qquad (1)$$

$$x_n = s(\omega)e^{-j\omega n\frac{D}{c}\cos(\Theta)} \qquad (2)$$

D = sensor spacing, c = speed of sound, Θ = angle to the source, and s = farfield source. Beamformers are classified into essentially 4 categories; delay and sum, oversteered, superdirective, and adaptive. Listed in order of increasing performance and complexity, the categories are defined by the general form of the weights and the method used in obtaining them. A detailed description for each category can be found in [4] while performance across categories is investigated by Kates and Weiss in [6].

During the first year of the DFBF project a significant amount of information was obtained by simulating and comparing different array geometries. Simulations included factors such as the number of sensors, sensor spacing, linear and planar arrangements, single and multiband filter structures. The geometries were evaluated in terms of their ability to meet specific performance criteria such as 3dB beamwidth, directivity versus frequency, and overall frequency response. Additionally, arrays were evaluated in terms of their computational complexity because the signal processing structure of the beamformer would ultimately determine the system's ability to maintain the required data throughput in real-time.

Figure 2 is a block diagram showing the structure of the beamformer as implemented in the DFBF microphone. The beamforming system generates four directional beams, each covering approximately 90 degrees in the horizontal plane. Only one beam is active at any given time. Rotation control for the system (not shown) is performed by an interrupt service routine (ISR) which multiplexes digitized audio from the appropriate codec channels into the frame buffer. If the desired direction is along the x_0 end of the array (defined as North), the ISR scans the codec channels in order from x_0 to x_7. Alternatively, if the desired direction is at the x_7 end (South), the ISR scans the codec channels in the reverse order, x_7 to x_0. Controlling beam rotation in this way requires only a single BF procedure, or thread, to process the audio. Additionally, since the audio from all microphones is highly correlated, this structure provides a seamless, artifact free rotation of the beam. In other words, there is no switching noise when transitioning between directions. Rotation of the beam is one of two adaptive procedures performed by the DFBF software.

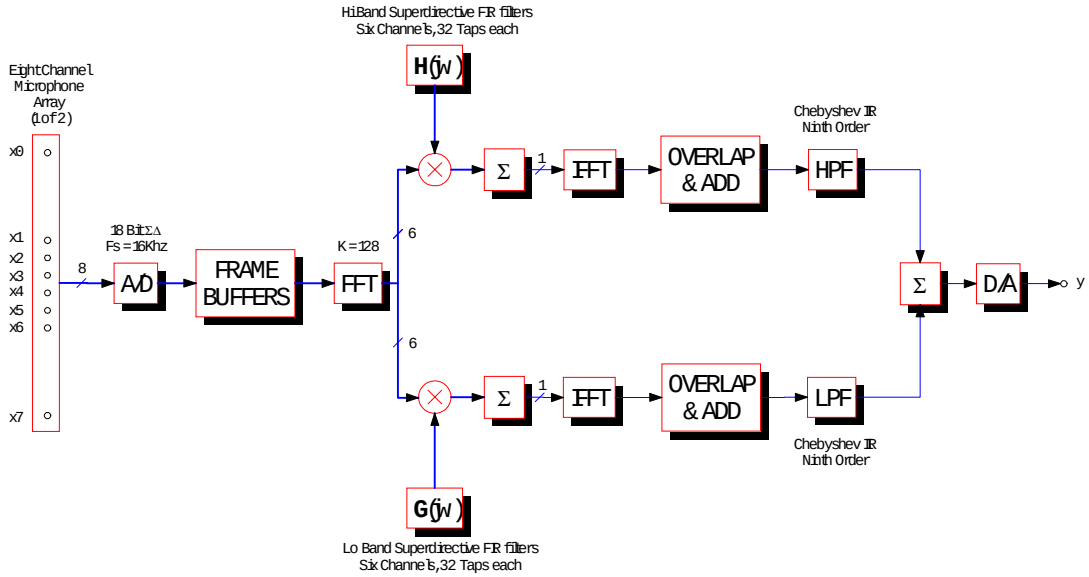
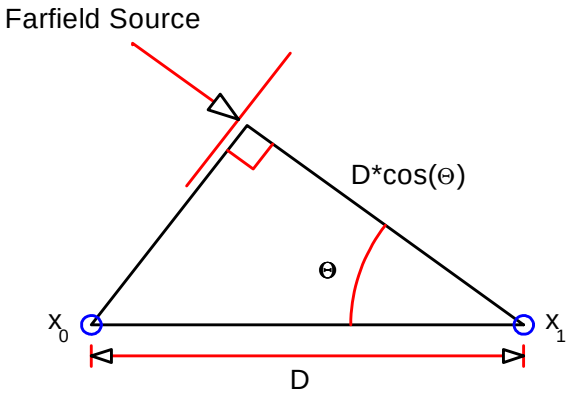


Figure 2. Beamformer Subsystem – Dual Band Acoustic Array

Superdirective filters for the array consist of six high band finite impulse response (FIR) filters, and six low band FIR filters. These filters are implemented in the frequency domain using fast convolution and an overlap and add process. Their purpose is to apply the proper phase delays to individual microphones signals so that a signal arriving along one end of the array is maximized. The FIR filter coefficients are stored as complex arrays. Beamformers often suffer from reduced directivity at low frequencies. For this reason, the eight channel array is partitioned into two sub-arrays. The crossover frequency between the sub-arrays is 1500Hz. Six microphones, x_1-x_6, are used as sources for the high band, while six other microphones, x_0,x_1,x_3,x_4,x_6,x_7 spaced further from each other, are used as sources for the low frequency band. Data for the twelve filters is processed and then the six outputs from each band are summed together. While this structure increases the number of superdirective filters from eight to twelve, it dramatically increases the directivity below the crossover frequency. The high and low band signals are transformed back into the time domain, passed through two infinite impulse response (IIR) crossover filters and summed to yield the output. The output is buffered and streamed to one D/A channel in the codec by the same ISR responsible for the inbound A/D data.

4. Direction Finding Subsystem

Sound source localization is a problem found in many areas of research such as underwater acoustics, ultrasound, and audio acoustics. Strictly speaking, source localization requires an estimation of two variables: relative bearing and either range or distance [9]. While both variables are necessary to specifiy the location of a source in 2D space, the DFBF system uses a simplified, bearing only estimate. Distance to the source offers little information in terms of controlling beam rotation, assuming at least some minimum distance from the array to the talker, or source. The diagram of Figure 4 assumes a single source residing in the extreme farfield so that the wave propagating across the array is effectively a plane wave. Using two sensors separated by a distance, D, and considering the assumptions mentioned previously, locating the angle to the source is reduced to the trigonometric relationship of (4), assuming an accurate estimate for the time delay, Δt, can be found. The parameter c is the velocity of sound in air. Since (3) relies on the cosine, there is an ambiguity for angles between +/-(π/2) radians. That is, for a given time delay, there is no way to determine if the angle to the source is +Θ or -Θ. This ambiguity is removed in the DF system by using four sensors located at the ends of both arrays. The polarities of the two $\cos^{-1}()$ operations are used to determine the quadrant of the expected angle, Θ.



$$\Delta t = (D/c)\cos(\Theta) \qquad (3)$$

$$\Theta = \cos^{-1}(\Delta tc/D) \qquad (4)$$

Figure 4. Time Delay Estimation for Plane Wave Propagation

While many methods exist for estimating time delay, most rely on one form or another of the cross-correlation function, which is in itself a computationally burdensome operation, particularly when calculated in the time domain. An efficient method, demonstrated in [2], calculates the generalized cross-correlation (GCC) function by estimating cross-spectral density, in the frequency domain, and then taking the inverse Fourier transform to obtain the GCC. This approach is commonly used in real time DSP applications since it requires far fewer computations

to perform complex multiplication in the frequency domain than it does to perform correlation (or convolution) in the time domain. The time delay estimate is obtained by determining the location along the time axis for which the GCC is maximum. Estimating the angle to the source is then the result of (4). It is important to note that the relationship between cross-spectral density and cross-correlation can only be applied to stationary random processes. In order to use this procedure on speech signals, which are non-stationary, it is necessary to limit the time frame over which the analysis is performed. It is often assumed that the speech spectrum will not vary significantly over a time period of 20-30 milliseconds. The GCC function, given by (5), is an efficient frequency domain approach to time delay estimation (TDE) problem.

$$Rx_0x_1(\tau) = \frac{1}{2\pi}\int W(\omega)X_0(\omega)X_1^*(\omega)e^{j\omega t}d\omega \qquad (5)$$

In equation (5) $X_0(\omega)X_1(\omega)$ is the cross-spectral density, * denotes complex conjugation, and $W(\omega)$ is a weighting function specific to a given application. The DFBF direction finder, shown in Figure 5, uses the 'Pitch Based Time Delay Estimation' method, described in [9], to derive this weighting function. While other, less complicated, weighting functions exist, this algorithm gives the DF component an inherent resistance to non-speech sounds and allows for very robust operation in the presence of background noise and reverberation.

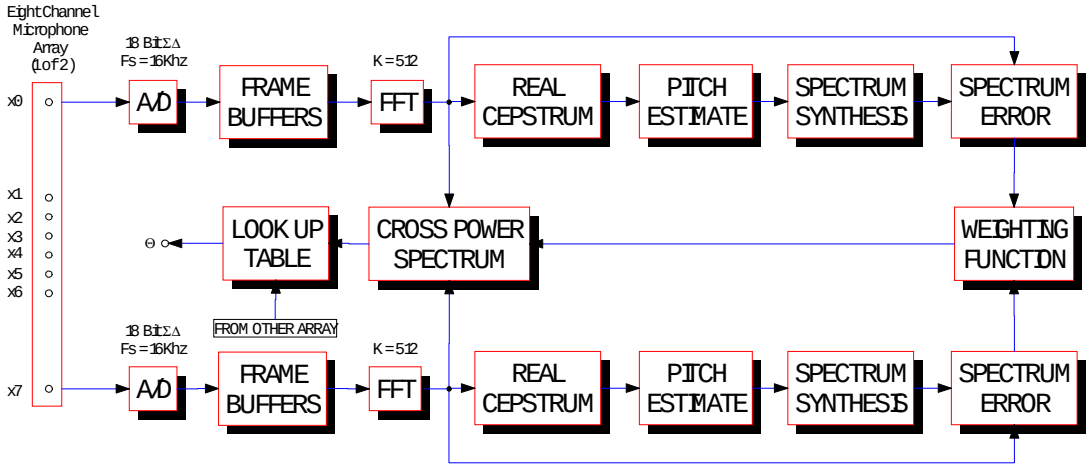


Figure 5. Direction Finding Subsystem – Pitch Based Time Delay

Cross-spectral density is a straight forward operation while generating the weighting function is a somewhat more complex process. The 'Pitch Based Time Delay' algorithm uses the periodic features found in voiced speech to estimate which spectral regions between the input signals are highly correlated. These correlated regions are emphasized in (5) while the uncorrelated regions, which perhaps contain noise or have been corrupted by reverberation, are ignored. Referring to the block diagram of Figure 5, the DF component generates the weighting function by first estimating the pitch, or fundamental frequency (F_o) of the speech. The F_o estimate is obtained by calculating the real cepstrum of a windowed speech sequence. F_o is used to generate a speech spectrum estimate by calculating the Fourier transform of a periodic impulse train. The original speech spectrum is then compared to the spectrum estimate and the error is used to generate the final weighting function, $W(\omega)$. Figure 5 shows only one half of the DF subsystem. An identical process is performed using the two microphones, x_8 and x_{15}, located at the ends of the other array. As mentioned previously, the additional pair of microphones is used to remove any ambiguity in the direction estimate. Time delay estimates from each microphone pair are

evaluated in terms of their polarity and magnitude, and if both fall within limits defined by the geometry, an updated direction is posted to the rotation control.

5. Real-Time Software

Time sequencing of the two processes (DF and BF) is performed using a finite state machine (FSM) since no real time operating system (RTOS) was available for this project. Although the DF component is more computationally demanding than the BF, it is the BF which is the primary audio data path, and as such must retain the highest priority in terms of DSP resource allocation.

Audio for the system is sampled at a rate of 16KHz. All signal processing, except in the ISR, is frame based where frames for the BF are 97 samples and those for the DF are 485 samples (6.0625 msec and 30.3125 msec), respectively. The frames are double buffered, or managed in a ping-pong fashion, so that processing can be performed in one half while data is being transferred into the other half. All BF data is zero padded and transformed with 128 point FFTs, which incidentally are TMS320C44 assembly language functions provided by Texas Instruments. The 97 sample sequences and 32 tap superdirective FIRs provide a 128 point, non-overlapping, output window after the filters have been applied. The resulting BF process has a througput latency of effectively one buffer, or approximately 6 msec, which is virtually unnoticed by most listeners. The entire BF process (12 frequency domain FIRs, 2 time domain IIRs) requires slightly less than 2.5 msec of the 6 msec budget defined by the BF frame size. This provides the DF subsystem 3.5 msec in which to perform the source localization algorithm.

The DF subsystem is composed of effectively four identical channels, each of which must generate a spectrum estimate for a 30 msec audio frame. The algorithm makes extensive use of internal (on chip) DSP memory and hand optimized assembly functions to reduce computation time. The overall process takes approximately 12 msec to obtain a direction estimate. Intuitively, it should take 4 (12msec(total)/3msec(alloc/frame)) frames to compute the DF process, however this assumes a fine granularity between the functions that make up the overall process. Unfortunately, this is not the case, and in fact this is where the system would greatly benefit from a RTOS. The DF process is segmented into 15 time slices by a FSM that works around the BF algorithm. This segmentation, and the time required to fill the 485 sample buffer, yields a direction estimate approximately every 75 msec. Fortunately, since speech, or more importantly conversation, is a relatively slowly changing process, the 75 millisecond computation time does not adversely affect the overall performance.

The DF and BF interface, as mentioned previously, is the ISR responsible for transferring audio. Every 6 msec the BF calculates the array output while every 75 msec the DF posts a new source direction, if appropriate. The two components work in concert to provide a high quality audio output which can be used in the applications mentioned previously.

6. Conclusion

Lexington Center/School for the Deaf currently has two DFBF conference microphone prototypes. The first prototype is fully operational and is being used as an in-house evaluation platform. The second prototype is being integrated now and will be used as a field evaluation system during the Fall of 2000. Measurements being performed on the in-house unit are intended to evaluate the quantitative performance. An on-site anechoic chamber will be used to measure directional performance using various acoustic and audio software packages, and an automated polar pattern measurement system designed for this project. The field prototype will be used in subjective experiments designed to evaluate the system in its intended environment. These experiments will determine the benefit the system provides in terms of reducing the effects of background noise and/or reverberation. More specifically, they are designed to evaluate any improvement in speech intelligilility that would result from using the DFBF conference microphone system.

7. Acknowledgements

The work presented here was funded by Field Initated Research Grant CDFA#84.133G received from the National Institute on Disability and Rehabilitation Research (NIDRR).

8. References

[1] Bakke, M., Levitt, H. and Steele, M. (1996). A Direction-Finding, Beam-Forming (DF-BF) Conference Microphone System. NIDDR Grant Application, CDFA#84.133G.

[2] Brandstein, M.S., Adcock, J.E, Silverman, H.F. (1995). A Practical Time-Delay Estimator for Localizing Speech Sources with an Array. Brown University, Providence RI.

[3] Brandstein, M.S. (1997). A Pitch-Based Approach to Time-Delay Estimation of Reverberant Speech. Harvard University, Cambridge, MA.

[4] DeBrunner, V.E., McKinney, E.D. (1995). A Directional Adaptive Least-Mean-Square Acoustic Array for Hearing Aid Enhancements. J. Acoust. Soc. Am., 98(1), 437-444.

[5] Houtgast, T. and Steeneken, H.J.M. (1985). A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77(3), 1069-1077.

[6] Kates, J.M., Weiss, R.W. (1995). A comparison of hearing-aid array-processing techniques. J. Acoust. Soc. Am. 99(5), 3138-3148.

[7] LaPlante, M.P., Hendershot, G.E. and Moss, A.J. (1992). Assistive technology devices and home accessibility feature: Prevalence, payment, need and trends. Advance Data, Number 127, Atlanta: Vital ahd Health Statistics of the Centers for Disease Control and National Center for Health Statistics.

[8] Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. J. Acoust. Soc. Am. 63(2), 533-549.

[9] Quazi, A.H. (1981). An Overview on the Time Delay Esimate in Active and Passive Systems for Target Localization. IEEE Transations on Acoustics, Speech, and Signal Processing. 29(3), 527-533.

[10] Schein, J.D. and Delk, M.T. (1974). The Deaf Population of the United States. Silver Spring, MD: National Association of the Deaf.