Sound Source Localization in Domestic Environment

Xuehai Bian, James M. Rehg, Gregory D. Abowd
College of Computing, GVU Center,
Georgia Institute of Technology
Atlanta, GA 30332-0280 USA
{bxh|rehg|abowd}@cc.gatech.edu

Abstract

Sound source localization strategies can be traced back to radar and sonar localization systems. In the report, we will review the main challenges of sound source, especially talker, localization problem and current major strategies. We proposed a practical peak-weighted PHAT TDOA method to find reliable source location in the Awarehome, which is a residential lab in Georgia Tech. Finally, we discuss the possible application scenarios and propose future direction of our work.

Keywords:
Sound Source Localization; PHAT; A/V fusion;

1 Introduction

Sound Source Localization system (SSL) is to determine the location of audio sources based on the audio signals received by an array of microphones at different positions in the environment.

Recently, microphone arrays are used in the enhancement of SNR for speech signal[1], sound source localization[7][9][23][24], echo removal [2]. They are also applied in speech recognition [3], hearing aids [4]. Portable testbed which consists of 8 microphones is reported to have been built for different signal processing tasks in undergraduate course projects [5]. The future of microphone array will be mainly focused on various applications, such as voice input in the automobile, desktop PC, hearing aids, tele-conferencing [6]. However, different applications have different requirements for microphone arrays, such as affordable cost, smaller size, less computation requirement, robust algorithm, higher accuracy, noise and echo cancellation, etc.

Many researchers have summarized challenges in sound source localization with microphone array [7][8]. Based on our experiences, the main challenges for sound source localization lie in five aspects. 1. How to enhance signal noise ratio? The background noise can be quite complex and varies in different environments. In home environment, the noise might come from AC, refrigerator, TV, fans, etc. 2. The reverberation in a house room (Characterized by reverberation time) is hard to model because of multi-paths between the sound sources and receptive microphones in the environment. Every room might have different geometric deployment and walls with different acoustic properties. The furniture in the room will not be fixed in a location. 3. The speech signal is a broadband signal. Many narrowband processing methods applied in radar system, sonar system will fail here. Different categories of sound sources have different characteristics. In addition, the volume of speech might vary greatly from whispering to shouting. 4. The sound sources are usually intermittent and non-stationary. This makes it hard to use adaptive filter that needs long converging time. And if the sound source is moving quickly, the point sound source assumption can not work any more. 5. To make it more complex, there might be more than one sound sources existing in the environment. All these challenges make it a difficult task to track sound sources in real complex domestic environments. As it is difficult to model these different environments, all the current SSL systems are making assumptions about their environments or sound sources. When such assumptions are not satisfied, the algorithm will fail[9].

Sound source localization has its distinguished advantages over other localization technologies, like visual tracking, active-badge system, RFID location system etc. In some environments, such as the restroom, bedroom etc, the audio capturing is not as intrusive as the camera system. And sound source location will only record the location of the sound source and thus decrease the privacy concerns for the users.

SSL can provide 3D location information. In some application the height of the sound is very important. Sound Source Localization can easily recognize some sound-specific objects such as TV, conversation, putting dishes in the dishwasher etc. Above all, the data bandwidth processed for SSL is relatively lower. Typically, one 320*160 camera with 15 frame per second needs to process 2.3Mbyte per second, while for high resolution sampling of the stereo audio it only needs to process 44.1K *2 = 88.2kByte per second. Compared to active-badge system, SSL does not require the users to wear any devices.

The core algorithms of sound source localization can be traced back to earlier active radar and sonar localization system. In these active systems, objects send out preset signal to the intended objects and compare the echo signal in order to compute the location of the object. In passive systems, the system only receives signal from targets, which is usually moving. By wearing some device that generates preset signals, the receiver can compute the location with high accuracy because of high SNR in a very narrow frequency range. However, for the sound sources in a domestic environment, the home-made sound are not preset signals and thus is a challenging problem.

The goal of the technical report is to summarize the current SSL algorithms, including two-step TDOA, high-resolution spectral estimation, and steered beamforming methods. After the summarization, we will describe the hardware platform we build in our lab that captures the real room audio and software that access and visualize the captured data. The scalable sound capturing platform can capture the audio data up to 32 channels. Upon this, we provide a platform that can test different SSL algorithms with C++ or Matlab in a real home environment.

We proposed a peak-weighted full correlated PHAT-GCC to compute the sound source location, which is equivalent to SRP-PHAT method. Our localization accuracy is fairly good with real time processing in one room of the residential lab, where we set up our system.

The structure of the paper is organized as following. In part 2, we will review current three SSL algorithms. In part 3, we introduce the algorithm used in our system. Then system hardware and software structure are discussed in part 4. Finally, some interesting applications are discussed.

2 Current SSL Algorithms and Systems

The results of SSL might be affected by four factors: (1)The number and content of every sound sources as compared with the environment noises; (2)The acoustic properties of the environment, such as reverberation; (3) Quantity, quality and geometric deployment of receptive microphone arrays; (4) Algorithms.

We will divide the current algorithms into three main categories, which is quite similar to the categories discussed in [7]. (1) Time delay of arrival. The locating process is divided into two steps: Computing time delay estimation for each pair of microphones and searching the location of sound sources. (2) Steered beamforming power. In this method the responses of arrays are normally filtered, weighted and summed at different possible locations. The position that has the maximum power is assumed to be final sound source location. (3)High resolution Spectral Estimation based on the spatiospectral correlation matrix, which is computation intensive but this strategy is able to detect multiple sound sources.

2.1 Time Delay Of Arrival(TDOA)

TDOA or Time of Delay Estimation (TDE) is the most explored method among three strategies. TDOA strategy has been successfully applied to radar, sonar systems, where the signal band is narrow and SNR is high. However, for talker localization problem, it is usually inside a house where the background is noisy and the signal is usually considered as broadband signal. The reverberation is also much more complex in the room environment than in the sky or sea. Many possible supplemental filtering or post-processing technologies, which is specific in certain environment, need to be applied here.

In TDOA strategy, accurate computation of delay is the basis for final source location with high accuracy. In the first step, time delay of arrival is estimated for each pair of microphones. Traditionally, the general cross correlation between two signals is used in time domain [10].

Signal received at microphone i:

$$x_i(t) = a_i s(t - t_i) + n_i(t) \qquad (1)$$

where:

$s_i(t - t_i)$ is the delay signal;

$n_i(t)$ is the noise;

a_i is the attenuation factor for microphone i.

General Cross Correlation between channel 1 and channel 2 is :

$$y_{12}(t) = \int_{-\infty}^{\infty} x_1(t+t)x_2(t)dt \qquad (2)$$

To reduce the computation time, the cross correlation is usually implemented in frequency domain using equation 3, because FFT can save much computation time.

GCC in Freq domain:

$$R_{x_1x_2}(t) = \int_{-\infty}^{\infty} G_{x_1x_2}(f)e^{j2pft}df$$

$$= \int_{-\infty}^{\infty} X_1(f) \bullet Conj\ (X_2(f))e^{j2pft}df$$

$$\ldots\ldots\ldots(3)$$

where *Conj* is the complex conjugation function.*X*$_1$*(f)* and *X*$_2$*(f)* are the Fourier transform of x$_1$(t) and x$_2$(t).

However, because of noise and reverberation in the environment, some weight functions are applied to enhance the quality of the estimation, such as Phase Transform(PHAT), Roth Processor, ML, Eckart, etc[10][11][12]. The general cross correlation with the filter, $\Psi(f)$, will be expressed as:

$$\hat{R}_{x_1x_2}(t) = \int_{-\infty}^{\infty} \Psi(f) \cdot \hat{G}_{x_1x_2}(f)e^{j2pft}df$$

$$= \int_{-\infty}^{\infty} \Psi_1(f)X_1(f)Conj(\Psi_2(f)X_2(f))e^{j2pft}df$$

$$\ldots(4)$$

The Phat magnitude weighting function is:

$$\Psi(f) = \frac{1}{|X_1(f) \bullet conj(X_2(f))|} \qquad \ldots(5)$$

Roth autocorrelation weighting function is:

$$\Psi(f) = \frac{1}{X_1(f) \bullet conj(X_1(f))} \qquad \ldots(6)$$

SCOT filter is:

$$\Psi(f) = \frac{1}{\sqrt{X_1(f) \bullet conj(X_1(f)) \bullet X_2(f) \bullet conj(X_2(f))}}$$

$$\ldots(7)$$

Another proposed method to compute TDOA is to use time adaptive filter[13][14][15]. It assumes the signal to be the convolution of pulse response for a particular microphone and source. The noises in different microphone are independent with each other.

$$x_i(t) = h_i(t) * s_i(t) + n_i(t) \qquad \ldots(8)$$

An adaptive filter is suggested to get the direct path between microphones and can cancel reverberation in certain amount of time. It needs about 250ms converging time, which makes it hard to be applied in fast moving sound sources. From our experience of implementing it, it depends very much one how to select the starting point and it is hard to converge.

The second step of TDOA strategy is to find the location according to the delays computed between specific microphones, whose positions are already known. One delay-estimation will make the location subject to one-half of the hyperboloid of two sheets expressed in equation (9). If the sound sources are far away from the sensors, we can assume it to be a cone in the search space, which will make the search space simple.

$$TDOA_{12} = \frac{|L - L_1| - |L - L_2|}{C_{sound}} \qquad \ldots(9)$$

where: L is potential sound source location; L1,L2 is the location of Microphone 1 and 2. C$_{sound}$ *is the broadcasting speed in the air, which is 340 m/s.*

Theoretically, three pairs of Microphones are enough to determine the location of sound source[16]. However, more TDOA data will help reduce the error of estimation. Many strategies can be used to search for the sound source location in potential space. Maximum likelihood method assumes the error to be Gaussian distribution around the true location. [17]. But it is difficult to verify whether it is optimal in discrete time processing [13].

The least mean square function can be used in evaluating potential locations. The goal is to find a location point that minimizes the sum of square of difference between expected TDOA and measured TDOA. However, this method ignores the possible different errors of the delay in different pairs. We provide a weighted TDOA LMS method which will be discussed later.

Different search strategies can be used here. Newton-Raphson, Guass-Newton method, steepest descent algorithm are among possible choices.

The 2-step TDOA strategy is usually quick enough for real-time processing. The disadvantage is that it makes a premature decision on an intermediate TDOA in the first step, thus throwing away useful information. A better approach recently proposed would preserve all the intermediate information to the end and make an informed decision at the very last step [12]. Conceptually, this approach is similar to steered beamforming that finds the point in the space which yields maximum energy. But they differ in theoretical merits and algorithm complexity.

The window size that is applied to generate the location is very important in decreasing noise and reverberation effect. Generally, the more window size is used the result should be more reliable. The typical reverberation time of a moderate reverberation room has a reverberation of 200 to 300ms. So the window size of 400ms, which is about 17640 samples for a 44.1kHz sampling frequency, is enough to remove the reverberation effect. However, the larger the window size, the more latency of the tracking it might be, which makes it useless in some tracking task that needs the response time to less than 20-30ms. And shorter latency algorithms are suitable for multi-sound source separation.

2.2 Steered Response Power (SRP)

In this strategy, the array of the microphones will be focused on various locations. It searches for possible locations and get the time of direct flight from the source location to the respective microphone.

$$L_{beam}(q) = \int_{t_0 - W}^{t_0 + w} [\sum_{i=1}^{N} x_i(t + t_i)]^2 \, dt \quad \ldots(10)$$

Equation (10) is a typical beamforming evaluation equation. It will delay and sum certain function of the signal at different locations to find the optimum location that makes the objective function reaches the maximum This is also called *focalization*. Generally speaking, it is a non-linear optimization problem. We can use Newton-Raphson algorithm or steepest descent algorithm that iterates during the procedure in order to find the location. However, in the SSL problem with beamforming strategy, the objective function might have many local maximums, which makes the results rely more on the initial point where the search starts. Some researcher proposes searching strategy to solve multi-local maximums problem [9].The computation of objective function and searching for maximum peak in the space are two main tasks in steered response power method that might be computation intensive.

2.3 High Resolution Spectral Estimation

Traditionally, the strategy is mainly applied for narrowband signals. It is based upon the spatiospectral correlation matrix which can be derived from the signals received by the microphone array. This category includes autoregressive modeling, minimum variance spectral estimation. Some of these methods are limited to far-field, which means the sensors are supposed to be far away from the sound source, and linear array situation, where the microphone array are deployed in a line. It also needs time to search in the whole space for maximum. The maximum is supposed to be a sharp maximum and can be easily found [9][18].

3 Energy and Peak-Weighted based TDOA

3.1 PHAT TDOA Signal Processing

TDOA is the most popular SSL methods. As we stated out before, it needs two steps in finding the sound source location. The direct cross correlation fails in finding the TDOA in our system. We attribute it to the noise and reverberation in our test environment.

3.2 Peak-weighted PHAT Algorithm

Our system computes the PHAT-GCC in the frequency domain. However, by checking the computed delay in the system, we find that around 15% of these data are corrupted when the sound source is in a fix position. By examining these wrong data, we find these errors are mainly caused by two reasons. The first the energy of sound is too low at that time. Another reason is when which you check the peak of the correlation, we find these peaks are not so significant. By saying a peak is significant, we mean there are some other very high peaks around the detected maximum peak.

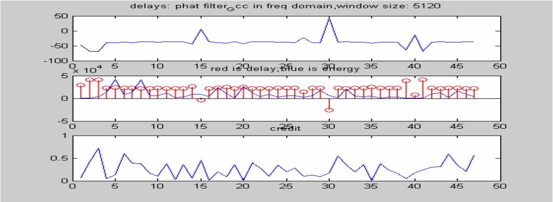

Figure 1 Typical TDOAs between two microphones compared with peak-weights and signal energy.

The upper sub-graph in Figure 1 shows the detected delays between two microphones along the time line in samples. In the middle sub-graph, the energy of one channel is displayed with the blue curve. The red stem lines are the delays (opposite directions) from first sub-graph. The lower sub-graph shows weights derived from peaks by equation(11). The window size we used is 5120 samples and the sampling rate is 44.1KHz.

Based upon above observation, we decide to use the ratio of the second peak with the maximum peak to convey the reliability of every time of delay. Specifically, we define the peak-weight to be:

$$\text{PeakWeight} = 1 - V_{\text{Second peak}} / V_{\text{Max Peak}} \quad \ldots(11)$$

The higher the peak-weight, the more weight we put on this microphone pair. Then the final goal evaluation function E is:

$$E = \sum_{i \in PossilbePairs} [W_i * (TDOA^{Exp}{}_i - TDOA_i)^2] \ldots(12)$$

Where W_i is the peak-weight and $TDOA^{Exp}{}_i$ is the expected time delay of arrival from a location. $TDOA_i$ is the calculated TDOA from captured audio data.

To find the location, we iterate according to the steepest gradient of equation (12) until minimum required error is satisfied. We find that the combination of peak-weight and energy of the signally has improved the resolution a lot in our environment. The results will be discussed in detail in section 4.2.

4 System Architecture

4.1 Hardware and sound capturing API

(1) Environment

The microphone array is installed in a guest room of the first floor in our residential lab, which is referred as Aware home [19].The dimension of the room is 360(L) by 305(W) by 272(H) cm. It has more furniture than a normal bedroom, including three large computer desks and many noisy computers. The fans of central AC are just outside of the room. Generally, it is a medium noisy environment as compared to normal house environment. See Figure 2.1 for the deployment of microphone array in the house.

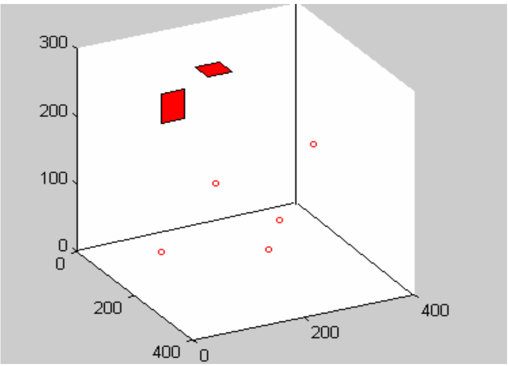

Figure 2. Deployment of microphone arrays in the room. Each red square represents a Quad.

(2) Geometric Design of Microphone Array

Four microphones are grouped together to form a microphone Quad. They are fixed at the vertex of a rectangle in a thin board. Assuming sampling rate to be 44.1Khz, during every sampling period, the sound can travel 340Kmm/44.1K= 7.71mm. In one typical Quad of our system, the edge of the square is 43.62cm, which corresponds to 80 samples(436.2mm*1.414/7.71mm = 80). Generally, we find the size of 10cm to 80cm is suitable for processing with less matching time. The microphone we are using is supercircuits-PA3 mini-microphone, Figure 3a. It is omni-directional and needs 12V pre-amplified power supply. The microphone can pick up the frequency range between 20-16KHz with 58db signal gain. It costs about $10 each, which is relatively much cheaper than some professional microphones and we found it is good enough in our environment.

Figure 3 (a) Omni-microphone. (b) Tango 24.

(3) Frontier Tango24 A/D converter.

The Tango A/D converter can convert electric sound signal from the microphone into the optical signal through ADAT interface, that can be connect to capturing device by standard optical fiber wire, Figure 3b. The advantage of transmitting through optical signal is to preserve the sound signal with least added transmission noise. The important properties that Tango 24 provides are:

- 8-ch A/D and D/A converter
- 44.1 or 48 kHz sampling rate
- Up to 24-bit quantization

(3) Frontier 16 channel Dakota sound capturing card.

Dakota PCI sound capturing card is a reliable sound card that reads signal from optical fiber wire through ADAT interface, which can be run in latest operating system, like Win2000 and WinXP. Normally, it can capture 16 input signals. Dakota is also extendable by using a Montana card which will extend it to 32 channels processing capability.

(4) Steinberg Audio Streaming Input Output(ASIO) development Kit.

One advantage of the Dakota is it can capture the 16 synchronized audio signals with the support of ASIO 2.0 programming API. ASIO2.0 is an industrial standard proposed by Steinberg software and hardware company in Germany. ASIO programming kit is a generally accepted multi-channel programming standard as compared with Microsoft's directSSK. If the signal is not synchronized between all the channels you want to pair up, you have to do it by writing your own code, which needs careful multi-threads programming and extensive CPU time in real-time processing according to our previous trial.

(5) Signal capturing and visualization.

Our system can capture the data and store it into an extended multi-channel wavefile format designed derived from standard wave file format. It visualizes different channels and allows zoom in/out both in amplitude and time to check the details of the signal. Figure 4 is the picture of captured data of 8 channels with different time of delays.

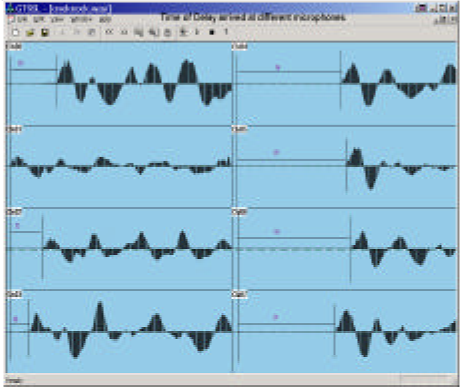

Figure 4 Visualized 8 channels audio signal.

4.2 Source localization system

Fig 5 is the sound source location history tracked by our system. The room is simulated by one front and two lateral walls, as well as a ground floor, with OpenGL. Every dot in the graph is a detected sound source with timestamp. The latest data are rendered with more red color and the older data are rendered with more green color.

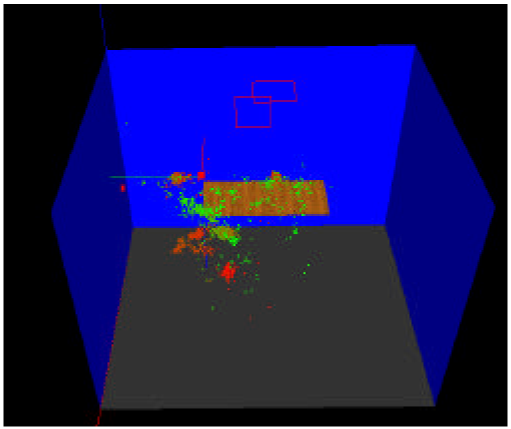

Figure 5 SSL history of half an hour conversation.

To test the effectiveness of the system, we place a fixed sound source, which is a speaker playing online NPR news, and test the result. The standard deviation varies with the source location. At these five different locations, we point the speaker to the direction of microphone arrays to enhance the SNR during our experiments. The SNR is around 5-20Db. The results are reported in Table 1.

Loc(cm)	Description	Num	Std Error (cm)
(58,221,83)	Desk	5365	(0.8, 1.0,1.7)
(220,313,179)	Right Back Top	3385	(2.1,2.0,1.4)
(264,15,84)	Left Back Desk	1453	(3.0, 3.0,2.2)
(161,262,11)	Ground	465	(4.4, 10.9, 14.6)
(287,215,104)	Lower Closet	3987	(9.0, 1.7,4.0)

Table 1. The standard errors of peak-weighted algorithm used in our system.

We can from the table that the standard errors vary with the location of sound sources and has larger standard error in those areas which have zero TDOA to many microphone pairs. We notice that in these areas, even small TDOA errors along corresponding axis would cause large relative error which consequently causes larger location estimation error in that axis. Another phenomenon is that, generally, for the sound location where it is very close to the two Quads, such as on the desk that is closest to the two Quads, the standard error is only 1-2 cms. However, for those locations which are far away from the two quads, it might be 10-14 centimeters. We suspect it is because in the area which is closer to the microphone array, the SNR is higher and thus the result is more accurate.

We are satisfied with the results in general, as the error is enough for us to distinguish between different loudest sound sources. In our experiment, we are using a Canon VCC3 PTZ camera which is driven by the sound source locator system. Most time it can exactly point toward the current speaker's mouth when there are some guests making conversation in our lab. Similar applications are also reported [25].

Another interesting property of our system is that it is easier to detect pulse sound, such as knocking the desk. It is also more responsive to woman's noise and less responsive to mundane sound.

5 Application of Sound Source Localization

An array of microphones bears more power than a single directional microphone. Microphone arrays are recently used in more and more systems, such as conference systems [20][22], acoustic surveillance, etc. In [6], the authors also provide many conceptual future applications with microphone array system.

However, the disadvantage of sound source tracking is also obvious: it will only track objects that produce sound. Basically, it is only a discrete location reporting system.

Beside the current, PTZ camera demo system, we are interested in building more applications with sound source localization. Below are some possible applications we will try to build with source localization abilities. We will categorize these applications according to the demands of sound processing technologies.

Scenario 1: Single sound source SSL

There is a map of the floor in the house, which is divided into small regions. The map will display the history of sound source events in corresponding area. The latest source, eg, in the past 20 seconds will be flashing with time showing freshness. This application can give an observer the impression of where sound activities mostly happens and frequency of the activity in a particular day. It can be fed into Digital Family Portrait[21] in Awarehome.

Scenario 2: Using SSL and speaker ID

Mr. Philips is at the porch of his home. He says "Home manager, I am back!" The door is automatically unlocked and says "Welcome home, Mr. Philips" according to his speech characteristics. The lamp system and display system will choose which lamps and displays to be turned on according to his footsteps or voice command.

Scenario 3: Using SSL, Speech recognition, Speaker ID

Light on: A says "Light On/Off", the light closest to his location is turned on. The intensity is tuned to different user's requirement.

"Message": Displays A's message in the closest display or voice message through the closes speaker.

Scenario 4: Using multi-talker SSL and speaker

ID- Meeting capturing system:

When there are more two talkers in the room, the large shared display is turned on. We might organize the audio data according to different speaker and point the camera to the current speaker that automatically captures the meeting.

All the above scenarios require robust sound source localization. Our future direction is to spend time in providing multi-talker tracking as well as building these interesting applications

6 Conclusions

In this report, we summarize the current main sound source location algorithms and introduce the platform we used in our lab

We find that in our current hardware deployment there are still many inevitable errors in time of delay calculation. We proposed our algorithm which uses peak-weighted goal function that detects sound source location in real time. Then some possible applications are discussed.

We are exploring the possibility of taking advantage of second peak, with particle filtering in order to get more reported sound source location data.

References

1. J. Bitzer, K. Simmer. "Superdirective microphone arrays", In M. S. Brandstein and D. B.Ward (eds.), Microphone Arrays: Signal Processing Techniques and Applications, Springer , 2001.

2. W. Kellermann. Integrating Acoustic Echo Cancellation with Adaptive Beamforming Microphone Arrays, Invited lecture at Forum Acusticum 99, Berlin, March 1999.

3. M. Omologo, M. Matassoni, P. Svaizer. "Speech recognition with microphone arrays", In M. S. Brandstein and D. B.Ward (eds.) , Microphone Arrays: Signal Processing Techniques and Applications, Springer Verlag, 2001.

4. Greenberg, Zurek, "Evaluation of an adaptive-beamforming method for hearing aids", J. Acoustical Society of America, vol. 91, pp. 1662-1676, 1992;

5. R. Moses and L C. Potter, "An Acoustic Array for undergraduate Instruction". Proc. of First Signal Processing Education Workshop (SPE2000), Hunt, TX, October 15-18, 2000

6. D. Compernolle, "Future Directions in Microphone Array Processing". In M. S. Brandstein and D. B.Ward (eds.), Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001. P389-394.

7. J. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments", PhD thesis, Brown University, May 2000.

8. S. Birchfield, D. Gillmor, "Acoustic source direction by hemisphere sampling", Proc. of ICASSP, 2001.

9.J. DiBiase, H. Silverman, and M. Brandstein, "Robust Localization in Reverberant Rooms, " In M. S. Brandstein and D. B.Ward (eds.) , Microphone Arrays: Signal Processing Techniques and Applications, Springer, 2001.

10.C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech, Signal Process. ASSP-24, P320-327, 1976.

11.Y. Rui and D. Florencio, "Time delay estimation in the presence of correlated noise and reverberation", Microsoft Research Tech Report, 2003-01.

12.Y. Rui and Dinei Florencio, "New direct approaches to robust sound source localization" , Proc. of IEEE ICME 2003 , Baltimore, MD, July 6-9

13.Y, Huang. "Real-time acoustic source localization with passive microphone arrays". Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 2001.

14.J. Benesty, G. W. Elko, and R. M. Mersereau. "An efficient linear-correction least-squares approach to source localization". Proc. IEEE ASSP Workshop Apps. Signal Processing Audio Acoustics, Oct. 2001, pp. 67-70

15.J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for real-time acoustic source localization system". Proc. IEEE ICASSP, vol. 2, pp. 937-940, Mar. 1999

16. M. Brandstein, J. Adcock, and H. Silverman, "A Closed-Form Method for Finding Source Locations from Microphone-Array Time-Delay Estimates," In Proceedings of ICASSP95, pages 3019-3022, 1995

17.W.R. Hahn and S.A.Tretter, "Optimum processing for delay-vector estimation in passive signal arrays", IEEE Trans. Inform. Theory, vol18,May 1973.

18. J. KRolik, "Focused wide-band array processing for spatial spectral estimation," in ASAAP, vol.2, Prentice Hall, 1991.

19.Awarehome. http://www.awarehome.gatech.edu/

20.Ross Cutler, Larry Davis. "Look who's talking: Speaker detection using video and audio correlation." IEEE International Conference on Multimedia and Expo, 2000.

21.Mynatt, E.D., Rowan, J., Craighill, S. and Jacobs, A. . "Digital family portraits: Providing peace of mind for extended family members." CHI 2001.Seattle, Washington: ACM Press, pp. 333-340.

22.R. Cutler, Y. Rui, et. al., "Distributed meetings: a meeting capture and broadcasting system", Proc. of ACM Multimedia, Dec. 2002, France.

23.M. Brandstein and H. Silverman, "A Practical Methodology for Speech Source Localization with Microphone Arrays," Computer, Speech, and Language, 11(2):91-126, April 1997.

24.J. Kleban, "Combined acoustic and visual processing for video conferencing systems", MS Thesis, The State University of New Jersey, Rutgers, 2000.

25.H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing", Proc. of ICASSP, 1997.