VOICE SOURCE LOCALIZATION FOR AUTOMATIC CAMERA POINTING SYSTEM IN VIDEOCONFERENCING

Hong Wang *Peter Chu*

PictureTel Corporation M/S 635, 100 Minuteman Road, Andover, MA 01810-1031

ABSTRACT

This paper describes the voice source localization algorithm used in the PictureTel automatic camera pointing system (*LimeLight*^{TM}, Dynamic Speech Locating Technology). The system uses an array of 46cm wide and 30cm high, which contains 4 microphones, and is mounted on top of the monitor. The three dimensional position of a sound source is calculated from the time delays of 4 pairs of microphones. In time delay estimation, the averaging of signal onsets of each frequency band is combined with phase correlation to reduce the influence of noise and reverberation. With this approach, it is possible to provide reliable three dimensional voice source localization by a small microphone array. Post processing based on a priori knowledge is also introduced to eliminate the influences of reflections from furniture such as tables. Results of speech source localization under real conference room conditions will be given. Some system related issues will also be discussed.

1. INTRODUCTION

In current videoconferencing systems, participants have to manually control the camera so that the far end can have a proper view of the near end talker(s). This is both burdensome and distracting to the users. In many situations, when the participants are unwilling to operate or are not familiar with camera operation, the far end will get a whole group shot during the entire meeting. This greatly reduces the intimacy of visual communications. Technologies such as infra-red subject tracking for speaker location requires the participants to wear extra accessories[1]. Tracking via video requires the user to point the camera at the object first and frame the object manually to tell the tracking algorithm where to track [2]. Compared to the above technologies, voice activated camera pointing is the most natural way of automatic camera control. It does not require any effort from the users, and consequently is strongly desirable in videoconferencing.

Sound source localization has been a research topic since the seventies. Different approaches have been proposed and investigated[3]. Phase correlation was proposed as an approach that does not cause spreading of the peak of the correlation function. However, the disadvantage of this method has been that when the SNR is low in a frequency band, the detection errors will be accentuated. M.Omologo[4] introduced summation of all frames corresponding to an acoustic event. This improved the SNR over white noise. In this paper, we introduce data pruning in addition to averaging across frames. A background noise estimator will give both the noise level and crosscorrelation of the noise. A speech onset detector is used to eliminate the influence of noise and reverberation. Crosscorrelation of noise are subtracted from the crosscorrelation of the speech onsets, and the onsets are averaged across frames. In this way, reliable three dimensional sound source detection is achieved by using a small microphone array of 4 microphones.

Realtime sound source localization systems were reported by M. Brandstein [5] and D.V.Rabinkin [6] using two panel microphone arrays mounted on the walls where each array contains 4 microphones. When using sound localization to point the camera in videoconferencing, it is very important that the relative position of the camera and the microphone array is well calibrated. Separate microphone array panels will require calibration in installation, and whenever the videoconferencing system is moved, re-calibration will be necessary.

By using the algorithm proposed in this paper, it is possible to build a small microphone array, and have the camera mounted on the same base with the array so that user installation is possible. Also, transportation of the videoconferencing system will not require re-calibration.

Besides noise and reverberation, one of the problems of sound localization in a videoconferencing room is the influence of specular reflection from the flat surfaces of the furniture such as a table. The crosscorrelation coefficient of the lag corresponding to the mirror image is sometimes greater than the real sound source. In order to eliminate the influence of this reflection, post processing is performed using a priori knowledge that in regular videoconferencing conditions, the angle of the reflection is always farther from perpendicular to the array than that of the original signal.

2. TIME DELAY ESTIMATION

Let $s(n)$ be the source signal, and $x_1(n)$, $x_2(n)$ be the signals received at the two sensors under additive noise conditions. The following relations hold assuming no multi-path distortions.

$$\begin{aligned} x_1(n) &= \alpha s(n-D) + n_1(n) \\ x_2(n) &= \beta s(n) + n_2(n) \end{aligned} \tag{1}$$

Where $n_1(n)$ and $n_2(n)$ represent noise signals received at the two sensors. The crosscorrelation of the signals received at the two sensors is

$$R_{x_1x_2}(\tau) = \alpha\beta R_{ss}(\tau - D) + R_{n_1n_2}(\tau), \tag{2}$$

assuming that the source signal is not correlated with the noise. D represents the delay of signal arrival between the two sensors. The cross power spectrum (The Fourier Transform of Eq(2)) is

$$G_{x_1x_2}(\omega) = \alpha\beta G_{ss}(\omega)e^{-j\omega D} + G_{n_1n_2}(\omega). \tag{3}$$

We assume that the background noise is stationary, but might be correlated (such as the noise from a ceiling fan

or an overhead projector, etc). Background noise $G_{n_1n_2}(\omega)$ is estimated for each frequency during a period when the source signal is not present. Subtracting the background noise, the cross spectrum becomes

$$G'_{x_1x_2}(\omega) = G_{x_1x_2}(\omega) - G_{n_1n_2}(\omega) = \alpha\beta G_{ss}(\omega)e^{-j\omega D}. \quad (4)$$

The normalized crosscorrelation, i.e. the Phase Transform [7] is therefore given by

$$\begin{aligned} \hat{R}_{x_1x_2}(\tau) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{G'_{x_1x_2}(\omega)}{|G'_{x_1x_2}(\omega)|}e^{j\omega\tau}d\omega \\ &= \delta(\tau - D) \end{aligned} \quad (5)$$

The time delay is estimated as the time lag that has the maximum cross-correlation coefficient

$$D = \text{argmax}_\tau \hat{R}_{x_1x_2}(\tau). \quad (6)$$

The reason that the above phase correlation approach also works well under reverberant conditions can be explained from the point of view of optimum detection. Assuming the noise signals received at both sensors (as shown in Eq(1)) have the same power spectrum $|N(\omega)|^2$, the optimum detection can be achieved by passing the signals through a whitening filter and calculating the cross-correlations as follows:

$$D = \text{argmax}_\tau \hat{R}_{x_1x_2}(\tau) = \text{IFFT}(\frac{G_{x_1x_2}(\omega)}{|N(\omega)|^2}) \quad (7)$$

When there are multi-path distortions, Eq(1) becomes

$$\begin{aligned} x_1(n) &= \alpha s(n-D) + h_1(n) * s(n) + n_1(n) \\ x_2(n) &= \beta s(n) + h_2(n) * s(n) + n_2(n) \end{aligned} \quad (8)$$

$h_1(n) * s(n)$ and $h_2(n) * s(n)$ represent the reverberations. If the reverberations are also considered to be noise, the overall noise component becomes

$$|N'(\omega)|^2 = |H(\omega)|^2|S(\omega)|^2 + |N(\omega)|^2. \quad (9)$$

Here we assume that the two reverberation transfer functions have the same power spectrum $|H(\omega)|^2$. The optimum detection of cross-spectrum in this case is

$$\hat{R}_{x_1x_2}(\tau) = \text{IFFT}(\frac{G_{x_1x_2}(\omega)}{|H(\omega)|^2|S(\omega)|^2 + |N(\omega)|^2}) \quad (10)$$

Assume that the reverberant energy is proportional to the direct energy, the following approximations can be obtained.

$$|H(\omega)|^2||S(\omega)|^2 = \gamma(|G_{x_1x_2}(\omega)| - |N(\omega)|^2) \quad (11)$$

$0 < \gamma < 1$ (for example, when we assume that the direct energy equals the reverberant energy, $\gamma = 0.5$). Then,

$$N'(\omega) = \gamma|G_{x_1x_2}(\omega)| + (1-\gamma)|N(\omega)|^2, \quad (12)$$

and the optimum delay estimation becomes

$$\begin{aligned} \hat{G}_{x_1x_2}(\omega) &= \frac{G_{x_1x_2}(\omega)}{\gamma|G_{x_1x_2}(\omega)|+(1-\gamma)|N(\omega)|^2} \\ D &= \text{argmax}_\tau \text{IFFT}(\hat{G}_{x_1x_2}(\omega)). \end{aligned} \quad (13)$$

A binary weighting is applied to the normalized cross-spectrum to eliminate the influence of noise.

$$\hat{G}_{x_1x_2}(\omega) = \begin{cases} \frac{G_{x_1x_2}(\omega)}{\gamma|G_{x_1x_2}(\omega)|+(1-\gamma)|N(\omega)|^2} \\ \simeq \frac{G_{x_1x_2}(\omega)}{\gamma|G_{x_1x_2}(\omega)|} & \textit{when signal} \gg \textit{noise} \\ 0 & \textit{otherwise} \end{cases} \quad (14)$$

The above equation is actually a modified case of the Phase Transform as shown in Eq(5), and it is also shown to be an optimum detection under multi-path distortion conditions.

The actual processing for calculating the time delay is as follows: digital samples from the two sensors are grouped into blocks (frames). The power spectra and cross spectrum are calculated for each frame. The magnitude of the power spectrum of one sensor is then compared with the background noise level (power spectrum and cross spectrum of background noise are estimated when there is no signals). If the frequency component of the power spectrum is above noise level by a certain threshold and is bigger compared to the signal power of previous frames by a certain threshold, the cross spectrum of these frequencies are chosen as signal onsets. The cross spectrum of signal onsets are then averaged over several frames and the cross spectrum of background noise is subtracted. The obtained cross spectrum is then normalized (as shown in Eq(14))to create $\hat{G}_{x_1x_2}(\omega)$ (The normalized cross spectrum $\hat{G}_{x_1x_2}(\omega)$ of the frequencies that do not have onsets are set to 0). Then inverse Fourier Transform is performed on the averaged normalized cross spectrum, and the time delay is chosen as the lag that corresponds to the maximum of the normalized cross correlation function.

3. AZIMUTH AND DEPTH DETECTION

The microphone array used for source localization is shown in Figure 1. Microphones 1 and 2 are used for azimuth detection. Microphone pair 1,3 and 2,3 are used for the depth detection. Since all microphones are mounted on the same plane, there are front-back ambiguities. In regular videoconferencing applications, we assume that only speech sources facing the monitor are of interest. Sound from the back of the monitor is attenuated by using cardioid microphones. Time delays of microphone pair 1,3 and 2,3 are interpolated using parabolic interpolation. The depth is calculated by triangulation. Table 1 shows the distance measurement results in 6 positions. The conditions of the room are: size :($5.5m \times 7.1m \times 3.1m$); reverberation time :$400ms$. Total number of measurements in each position is 25. The sampling rate is 16kHz.

4. ELEVATION ANGLE DETECTION

Microphones 3 and 4 are used for elevation angle detection. When there is flat furniture such as a table between the talker and the microphone array, sometimes the mirror image has a greater crosscorrelation coefficient than the real sound source. Simply choosing the time lag which gives the maximum crosscorrelation coefficient might erroneously result in the elevation angle of the reflection of the source versus the source being chosen.

In a video conference, the camera is usually mounted on top of the monitor to keep the far end and near end eye contacts as natural as possible. In this case, it is usually true that the time lag corresponding to the reflection of the sound source is farther away from perpendicular to the array. Figure 2 shows a typical cross correlation function where the time lag corresponding to the reflection of the sound source has a greater cross correlation coefficient than the real sound source. To eliminate this problem, a threshold is introduced to group the peaks in the crosscorrelation function. Then the peak that is closer to zero lag is chosen as the real time lag. Most of the detection errors due to table reflections are eliminated by using this approach.

5. SYSTEM CONSIDERATIONS

Other issues besides accurate detection of speech source location also need to be considered in order to make the camera pointing system a usable product. Important issues include 1) Accurate speech detection. A speech detector is integrated to prevent the system from responding to non-speech signal. 2) Far end signal detection. Without far end signal detection, far end speech that comes out of the loud speaker might be detected as near end speech, and cause the camera to point at the loud speaker. 3) An intelligent camera control layer is necessary to make decisions of where and when to point the camera. For example, depending on the situation, sometimes it is desirable to show the person who is speaking, some other times showing a small group of people who are having a discussion is more appropriate than letting the camera follow each speaker. Since extraneous camera move causes discomfort in video conferencing for the far end viewers, it is very important to perform proper camera pointing with the least amount of camera move.

The PictureTel *LimeLight*^{TM}[8], shown in Figure 3, consists of the intelligent sound source localization system and the PictureTel PowerCam 100, a custom designed camera with a high-speed, highly accurate pan-tilt-zoom drive mechanism.

6. CONCLUSIONS

In this paper, we described the time delay estimation algorithm used in the PictureTel *LimeLight*^{TM}, the first commercially available voice activated automatic camera pointing system. The concepts of data pruning, data averaging and post processing combined with phase correlation make it possible to realize reliable three dimensional sound source localization using a small microphone array. A layer of user interface control is also designed on top of the sound source localization algorithm to make sure that the camera performs appropriate framing.

Table 1: Standard deviation of distance measurements (average of 25 measurements)

Distance (meters)	Standard deviation	Azimuth (degrees)	Elevation (degrees)	number of outliers
2.03	0.06	-26.1	-18.5	0
2.12	0.15	24.5	-16.3	0
2.76	0.20	-20.6	-11.2	0
3.37	0.24	33.4	-9.1	0
4.08	0.22	15.4	-9.1	0
4.39	0.38	14.9	-6.1	1

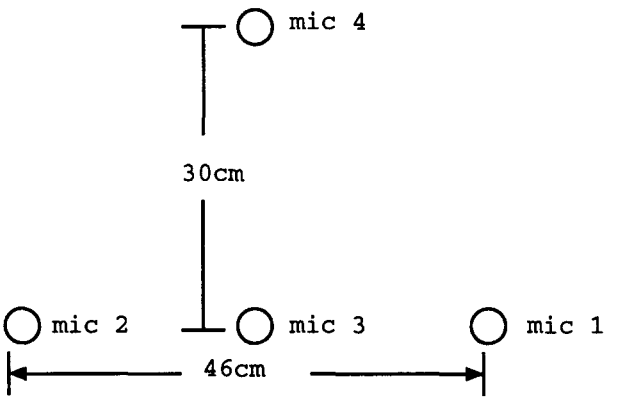

Figure 1: Microphone array for sound source localization

ACKNOWLEDGMENTS

The authors thank PictureTel research staff for their discussions, comments and support. We thank PictureTel human factors engineer for helping with user interface design and for conducting numerous user tests. We thank PictureTel cross product division for their help on hardware design and realtime implementation. Thanks also go to PictureTel volunteers for using our early prototypes in their business meetings and providing valuable feedbacks.

7. References

[1] *Camera tracking technology, CameraMan*^{TM} *by ParkerVision.*

[2] *Video tracking camera, EVI-D30, by SONY.*

[3] G. C. Carter, editor. *Coherence and Time Delay Estimation.* IEEE Press, 1993.

[4] M.Omologo and P.Svaizer. "Acoustic Source Location in Noisy and Reverberant Environment Using CSP Analysis". In *proceedings of the ICASSP IEEE*, pages 921–924, 1996.

[5] M.Brandstein and H. Silverman. "A Localization-Error-Based Method for Microphone-Array Design". In *proceedings of the ICASSP IEEE*, pages 901–904, 1996.

[6] D.V.Rabinkin, R.J.Ranomeron, A.Dahl, J.French and J.L.Flanagan. "A DSP Implementation of Source Location Using Microphone Arrays". In *3aEA1 proceedings of the JASA*, 1996.

[7] C.H.Knapp and G.C.Carter. "The Generalized Correlation Method for Estimation of Time Delay". In *IEEE Trans. ASSP*, volume 24, pages 320–327, 1976.

[8] *US Patent application pending.*

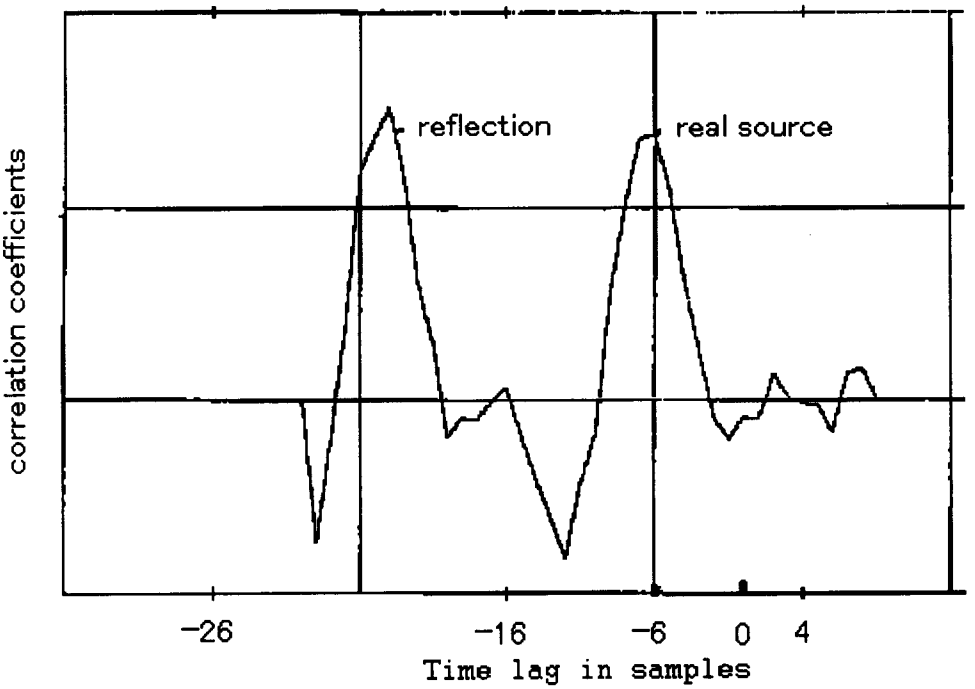

Figure 2: A typical crosscorrelation function with specular reflections

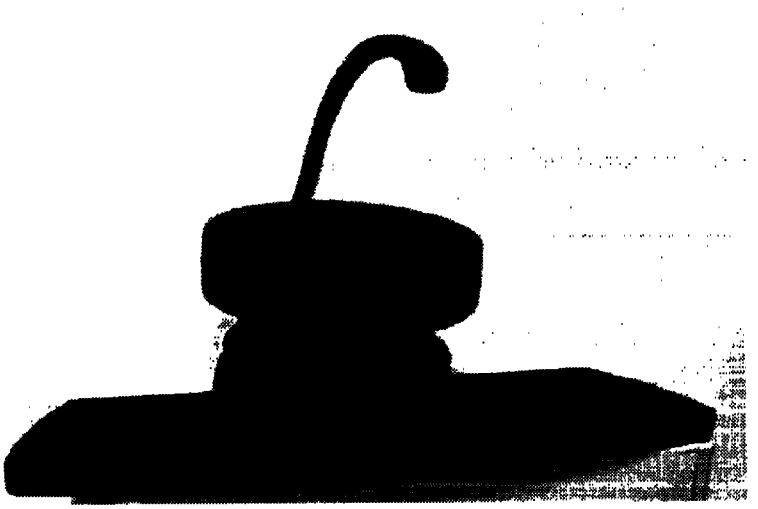

Figure 3: The PictureTel *LimeLight*^{TM}, Dynamic Speech Locating Technology