ENERGY-BASED POSITION ESTIMATION OF MICROPHONES AND SPEAKERS FOR AD HOC MICROPHONE ARRAYS

Minghua Chen, Zicheng Liu, Li-Wei He, Phil Chou, Zhengyou Zhang

Microsoft Research
One Microsoft Way, Redmond, Wa 98052
{minghuac,zliu,lhe,pachou,zhang}@microsoft.com

ABSTRACT

We present a novel energy-based algorithm to estimate the positions of microphones and speakers in an ad hoc microphone array setting. Compared to traditional time-of-flight based approaches, energy-based approach has the advantage that it does not require accurate time synchronization. This property is particularly useful for ad hoc microphone arrays because highly accurate synchronization across microphones may be difficult to obtain since these microphones usually belong to different devices. This new algorithm extends our previous energy-based position estimation algorithm [1] in that it does not assume the speakers are in the same positions as their corresponding microphones. In fact, our new algorithm estimates both the microphones and speakers simultaneously. Experiment results are shown to demonstrate its performance improvement over the previous approach in [1], and evaluate its robustness against time synchronization errors.

1. INTRODUCTION

How to use sensor arrays such as microphones and cameras to improve meeting experience has attracted a lot of interests in the research community and the industry in the past several years. More and more meeting rooms use dedicated devices with built-in microphone arrays such as Polycom's SoundStation [2] and Microsoft's RingCam [3]. Microphone arrays are used to enhance audio quality as well as to localize speakers for camera directing. Since the individual microphones in a microphone array device are time synchronized and the geometry of the microphones are known, techniques based on time delay of arrival are used to determine the location of the speaker, known as sound source localization (SSL). Once the sound source is localized, intelligent mixing or beamformer picks up the sound and outputs higher quality audio. Furthermore, the estimated speaker locations are used to direct the video camera to point to the speaker providing better visual experience to the remote meeting participants. Over the past decade, tremendous progress has been made on time delay estimation, microphone array beamforming, and sound source localization techniques. It is formidable to list all the references here.

In this paper, we address a different type of microphone array, called ad hoc microphone array. An example of such an ad hoc microphone array is a set of microphones built in the laptops which are brought in by meeting participants. As portable devices are becoming increasingly popular in collaborative environments, many people bring their laptops and PDAs to meeting rooms. Many of these devices are WiFi enabled and have built-in microphones. They can easily form a network in an ad hoc fashion. We would like to leverage such ad hoc microphone networks to improve meeting experience.

Compared to traditional microphone arrays, on one hand, ad hoc microphone arrays are spatially distributed and the microphones are in general closer to the meeting participants. This is helpful for both audio quality improvement and sound source localization. Even in a meeting room with a dedicated microphone array device, the ad hoc microphone array could potentially help the dedicated microphone array device to obtain better audio quality and more accurate sound source localization since the ad hoc microphones are closer to the speakers. On the other hand, ad hoc microphone arrays present many technical challenges including (1) Microphones are not synchronized; (2) The array geometry is unknown; (3) Microphones have different and unknown gains; and (4) Microphones have different signal to noise (SNR) ratios.

Lienhart et. al. [4] developed a system to synchronize the audio signals by having the individual microphone devices to send special synchronization signals over a dedicated link. Raykar et. al. [5] developed an algorithm to calibrate the positions of the microphones and loudspeakers by having each loudspeaker to play a coded chirp. Time of flight can be reliably estimated by matching the special chirp signals.

In a meeting room environment, having the devices to play special signals may be distracting to people especially in the middle of the meeting when a device changes position or a new device joins in the network. A more natural approach is to estimate the positions based on human speech signals.

Recently Liu et. al. [1] developed an energy-based technique to estimate the microphone positions and gains based on human speech signals. They used audio signal energy decays to estimate the distances between the devices. After obtaining the pairwise distance estimations, they used a multidimensional scaling technique to compute the coordinates.

One limitation of their technique is that it assumes each speaker is at the same position as the corresponding microphone. In this paper, we present a more general framework that does not have this limitation. With this new technique, we are able to estimate both the microphone and speaker positions simultaneously. We show experiment results to demonstrate its performance improvement over the previous approach in [1], and evaluate its robustness against time synchronization errors.

2. FORMULATION OF ENERGY-BASED POSITION ESTIMATION

Suppose there are m microphones and n speakers in a meeting room. We assume they are on the same 2-D plane. Let $z_i(t)$, $i = 1, ..., m$ denote the audio stream captured by the i-th microphone.

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

Let a_{ij} denote the average energy of the audio segment in $z_i(t)$ that corresponds to j-th person's speech. Let s_j denote the average energy of j-th person's original speech which is unknown. We sometimes call s_j the volume of speaker j. Let c_{ij} denote the attenuation of person j's speech when it reaches microphone i. Let m_i denote the gain of microphone i. In the absence of noise and observation error, a_{ij} is $m_i s_j c_{ij}$. Modelling noise and observation error as a zero mean Gaussian random variable in log domain, we express the noisy observation a_{ij} as follows,

$$\ln(a_{ij}) = \ln(m_i s_j c_{ij}) + \epsilon_{ij}, \tag{1}$$

where $\epsilon_{ij} \sim N(0, \sigma_{ij}^2)$. This model is similar to the lognormal receiving power model widely used in wireless communication [6].

Let d_{ij} denote the Euclidean distance between microphone i and speaker j. The relationship between c_{ij} and d_{ij} is in general modelled as $log_{10} c_{ij} = -\lambda log_{10} d_{ij}$ where λ is a parameter depending on the environment [7]. Since we are mainly interested in meeting room environment, we measured a typical meeting room in an office building and found that λ is approximately 1 [1]. Thus we assume $c_{ij} = \frac{1}{d_{ij}}$.

Let $P_i^m = (x_i^m, y_i^m)$ denote the position of i-th microphone, and $P_j^s = (x_j^s, y_j^s)$ denote the position of j-th speaker. Then Eqn. 1 can be expressed as

$$\ln(a_{ij}) = \ln\left(\frac{m_i s_j}{\sqrt{(x_i^m - x_j^s)^2 + (y_i^m - y_j^s)^2}}\right) + \epsilon_{ij}. \tag{2}$$

That is,

$$\begin{aligned} \ln(a_{ij}) = \quad & \ln(m_i) + \ln(s_j) \\ & -\tfrac{1}{2}\ln((x_i^m - x_j^s)^2 + (y_i^m - y_j^s)^2) + \epsilon_{ij}, \\ & i = 1, ..., m, j = 1, ..., n. \end{aligned} \tag{3}$$

We use maximum likelihood estimation to generate estimates for the positions of microphones and speakers. The maximum likelihood estimates are given by solving the following problem:

$$\begin{aligned} \arg\max_{m_i, x_i^m, y_i^m, s_j, x_j^s, y_j^s} & \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{1}{\sigma_{ij}^2}(\ln(a_{ij}) - \ln(m_i) \\ & - \ln(s_j) + \frac{1}{2}\ln((x_i^m - x_j^s)^2 + (y_i^m - y_j^s)^2)). \end{aligned} \tag{4}$$

There are $3(m+n)$ variables to estimate: $m_i, x_i^m, y_i^m, s_j, x_j^s, y_j^s$, $i = 1, ..., m$, $j = 1, ..., n$. Since the microphone and speaker positions can only be determined up to a global translation and rotation, and both microphone gains and speaker volumes can only be determined up to a scaling, the actual number of uncertainty to resolve for is $3(m+n)-5$. As the number of observations is mn, the optimization problem is well defined if $mn \geq 3(m+n) - 5$.

3. INITIALIZATION

Eqn. 4 is a nonlinear optimization problem with non-convex objective function, and has no close form solution. As such, we solve it by using a commonly used numerical optimization technique, Levenberg-Marquardt method [8]. Due to the non-convexity of the optimization problem, the numerical procedure often get stuck in a local minima if the initial guesses are too far away from the optimum solution.

In order to perform the maximum likelihood estimation, we need to measure a_{ij} and δ_{ij}, and generate good initial guesses. We will focus on generating initial guesses in this section, and will describe how to measure a_{ij} and δ_{ij} in the next section.

In order to obtain reasonable initial guesses, we first assume each speaker is at the same position as its corresponding microphone and uses the technique as described in [1] to estimate microphone gains, microphone positions, and speaker volumes. The correspondence between the speakers and microphones is determined heuristically based on SNR information. When the number of speakers is larger than the number of microphones, we cluster the speakers into groups so that the number of groups is the same as the number of microphones. Each group is then treated as a single speaker.

In the following, we briefly describe the procedures for estimating microphone gains, microphone positions and speaker volumes under the assumption that each speaker is as the same position as the corresponding microphone. For detailed derivations, the reader is referred to [1].

First, c_{ij} are estimated by the following equation

$$c_{ij} = \sqrt{\frac{a_{ij}a_{ji}}{a_{ii}a_{jj}}}. \tag{5}$$

Then we obtain the pairwise distances between the microphones via the relation $d_{ij} = \frac{1}{c_{ij}}$. Given pairwise distances d_{ij}, a Singular Value Decomposition (SVD) based Multidimensional Scaling (MDS) Technique [9] is applied to compute the coordinates of the microphones.

We estimate the relative microphone gains by

$$\frac{m_j}{m_i} = \frac{a_{ji}}{a_{ii}}\frac{1}{c_{ji}}. \tag{6}$$

Since the gains are equivalent with respect to scaling, we can generate one set of microphone gains by setting $m_1 = 1$ and use Eqn. 6 to obtain the rest m_i. Based on estimated c_{ij} and m_i, the speaker volumes s_j are estimated from the following equation:

$$s_j = \frac{a_{ij}}{m_i c_{ij}}. \tag{7}$$

4. MEASUREMENT OF AVERAGE ENERGY AND VARIANCE

To measure a_{ij} and σ_{ij}, we first perform short window Fourier transformation on $z_i(t)$. Let $Z_i(k, f)$ denote the transformed signal where k is the frequency band and f is the frame number. Denote $b(f)$ to be the total energy of frame f, that is, $b(f) = \sum_k |Z_i(k, f)|^2$. Assume person j's speech segment is from f_1 to f_2. Then $\ln(a_{ij})$ and σ_{ij} are computed as

$$\ln(a_{ij}) = \frac{1}{f_2 - f_1 + 1}\sum_{f=f_1}^{f_2} \ln(b(f)), \tag{8}$$

$$\sigma_{ij} = \sqrt{\frac{\sum_{f=f_1}^{f_2}(\ln(b(f)) - \ln(a_{ij}))^2}{f_2 - f_1}}. \tag{9}$$

5. EXPERIMENT RESULTS

We use two sets of actual audio recordings to evaluate the performance of our proposed algorithm. The recording was done in a

meeting room of an office building. The reverberation time of the room is approximately 250 milliseconds.

In the first set of experiments, we place 7 synchronized stand-alone microphones on a meeting room table to represent 7 laptops and capture 7 people's voice signals. These microphones have different gains and the gains are unknown. Each person was asked to speak a short sentence. Since the microphones are synchronized, the seven audio files can be perfectly aligned. Then speaker segmentation is performed by finding the segment of the highest SNR on each audio file. The details are omitted since this is not the focus of this paper. To obtain the ground truth of microphone positions, we used a ruler to measure the pairwise distances between the microphones. Since the stand-alone microphones are quite small, the distance measurement is quite accurate. After the pairwise distances are measured, we use the SVD-based MDS [9] technique to compute the 2D coordinates from the measured distances. The resulting coordinates are used as the ground truth to evaluate the performance of our algorithm. In Fig. 1, the points marked with square signs are the ground truth of the seven microphone positions. We did not measure the ground truth of human speakers because people tend to move a lot when they are speaking and it is difficult to reliably measure the mouth positions of human speakers.

We first run the SVD algorithm in [1] to generate estimates of microphone positions (as mentioned earlier, the speaker positions are assumed to be at the same positions as the corresponding microphones), as well as microphone gains and speaker volumes. We then run our proposed algorithm with the SVD results as initial guess. Since the positions can only be determined up to a global transformation (rotation, translation, and scale), we compute a global transformation to align the estimated positions with the ground truth positions. We define the estimation error as the average distance between the aligned points and the ground truth positions.

The results are shown in Fig. 1 where the blue dots are the ground truth microphone positions, the green dots are microphone positions estimated by our proposed algorithm, and the red dots are estimated speaker positions. For visual comparison, the results of the SVD algorithm are shown in Fig. 2.

There are two improvements as compared to the SVD result. First, the estimation error of the SVD algorithm is 0.26 meters, while the estimation error of our proposed algorithm is 0.15 meters. The relative error reduction is 42%, which is significant. Second, our proposed algorithm generates quite reasonable speaker positions, which cannot be obtained by the previous SVD algorithm.

To evaluate the robustness of our algorithm against the synchronization error, we artificially introduce synchronization errors by randomly shifting each audio channel. The amount of shifting is generated by a random generator with a uniform distribution on $[-T, T]$ where T is a parameter that controls the average amount of shift on the audio channels.

By choosing different T values, we generate data with different amount of time shift (synchronization error). The results are shown in Fig. 3. The horizontal axis is the average amount of time shift of the audio channels. The vertical axis is the estimation error. The blue curve is the estimation error of the SVD results. The green curve is the estimation error of the results from our proposed algorithm. We can see that both algorithms are quite robust to synchronization errors. The estimation increases very slowly when the average synchronization error is less than 500 milliseconds.

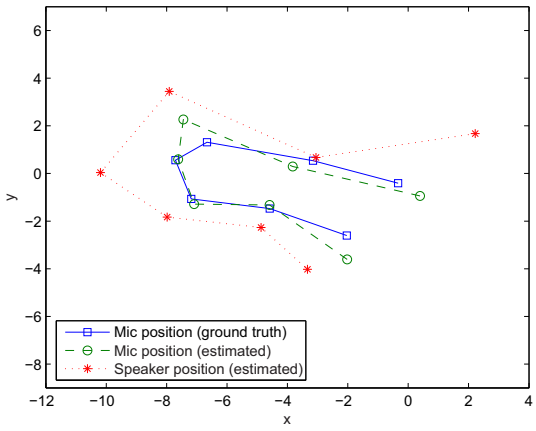

Figure 1: Estimated microphone and speaker positions on the synchronized data set. The blue curve is the ground truth microphone positions. The green curve and red curves are, respectively, the microphone and speaker positions estimated by our algorithm.

This shows the robustness of the energy-based approach against synchronization errors.

In addition, Fig. 3 shows that our proposed algorithm consistently improves the SVD results as long as the average synchronization error is no larger than 0.8 seconds. When the time shift is too big, SVD works better because of its regularization constraints on the speaker positions.

We have also tested our algorithm on the audio data recorded by laptops. We asked seven people each with a laptop sitting around a meeting table. The seven laptops have different brands and their microphone gains are unknown. Each person was asked to speak a short sentence.

The audio files recorded by the laptops are roughly aligned by detecting the first speech frame on each audio channel through simple thresholding. To obtain the ground truth, we measured the pairwise distances between the microphones on the laptops. The microphones are located by visual inspection. We then use Multi-dimensional Scaling technique to compute the 2D coordinates of the laptop microphone positions.

Fig. 4 shows the results of our algorithm. Again, the blue dots are the ground truth, the green dots are the microphone positions estimated by our algorithm and the red dots are the estimated speaker positions. The estimation error of our proposed algorithm is 0.21 meters. In comparison, the estimation error of the SVD algorithm is 0.26 meters. Our algorithm achieves 20% relative error reduction. Furthermore, the estimated speaker layout looks consistent with the actual speaker positions.

6. CONCLUSIONS

In this paper, we have presented a new energy-based algorithm to estimate the positions of both microphones and speakers in an ad hoc microphone array setting. In our approach, noise in receiving powers is modelled using lognormal distribution. We then perform a maximum likelihood estimation on the observed receiving powers to generate the results. Compared to our previous energy-based position estimation algorithm [1], the proposed algorithm has two

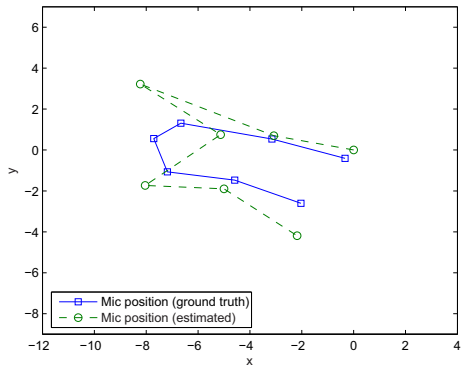

Figure 2: Microphone and speaker positions estimated by the previously published SVD-based algorithm [1]. The blue dots are the ground truth while the green dots are the microphone positions estimated by the SVD-based algorithm.

advantages. First, it does not assume the speakers are in the same positions as their corresponding microphones, and can estimate positions of microphones and speakers separately and simultaneously. Second, the new proposed scheme can generate more accurate estimates for microphone position than the previous algorithm. Results of actual experiment are shown to characterize the performance of the proposed algorithm, and evaluate its robustness against time synchronization errors.

7. REFERENCES

[1] Z. Liu, Z. Zhang, L.-W. He, and P. Chou, "Energy-based sound source localization and gain normalization for ad hoc microphone arrays," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, Honolulu, Hawaii*, 2007.

[2] "$http://www.polycom.com/products\_services$."

[3] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *ACM Multimedia*, 2002.

[4] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio processing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[5] V. C. Raykar, I. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, January 2005.

[6] G. L. Stuber, *Principle of mobile communication*, 2nd ed. Kluwer, 2001.

[7] T. Pham, B. M. Sadler, and H. Papadopoulos, "Energy-based source localization via ad-hoc acoustic sensor network," in *IEEE Workshop on Statistical Signal Processing*, 2003.

[8] N. Gershenfeld, *The nature of mathematical modeling*. Cambridge University Press, 1999.

[9] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, pp. 401–419, 1952.

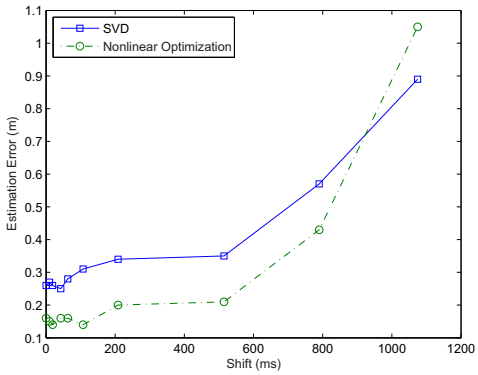

Figure 3: Performance evaluation when there are synchronization errors between audio channels. The horizontal axis is the average amount of time shift in milliseconds. The vertical axis is the estimation error in meters. The blue curve is the result from the previous SVD algorithm. The green curve is the result from our proposed algorithm.

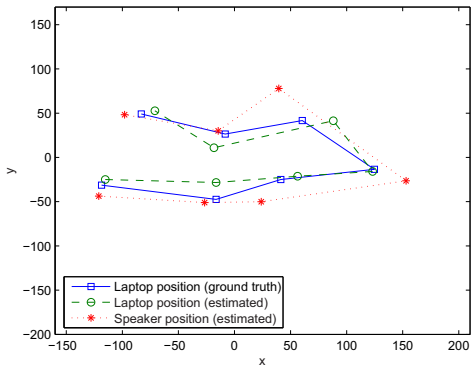

Figure 4: Results on the data set recorded by the laptops. The blue dots are the ground truth positions of the laptop microphones. The green and red dots are, respectively, the microphone and speaker positions estimated by our algorithm.