ENERGY-BASED SOUND SOURCE LOCALIZATION AND GAIN NORMALIZATION FOR AD HOC MICROPHONE ARRAYS

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

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

We present an energy-based technique to estimate both microphone and speaker/talker locations from an ad hoc network of microphones. An example of such ad hoc microphone network is a set of microphones built in the laptops that some meeting participants bring in a meeting room. Compared with traditional sound source localization approaches based on time of flight, our technique does not require accurate synchronization, and it does not require each laptop to emit special signals. We estimate the meeting participants' positions based on average energies of their speech signals. In addition, we present a technique, which is independent of the volumes of the speakers, to estimate the relative gains of the microphones. This is crucial to aggregate various audio channels from the ad hoc microphone network into a single stream for audio conferencing.

Keywords: Ad hoc microphone array, sound source location, meeting analysis

1. INTRODUCTION

Portable devices are becoming increasingly popular in collaborative environments such as in meetings, teleconferencing, and class rooms. Many people bring laptops and PDAs to meeting rooms and many of these devices are WiFi enabled and have built-in microphones. We are interested in how to leverage such ad hoc microphone networks to improve user experience in collaboration and communication.

Many audio conferencing in meeting rooms uses a special device with multiple microphones such as Polycom's SoundStation [1] and Microsoft's RingCam [2]. The device usually sits in the middle of the table. Because all microphones are synchronized and the geometry of the microphones is known, techniques such as time delay of arrival (TDOA) [3] can be used to determine the location of the speaker/talker, 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 than if a single microphone is used.

However, not every meeting room is equipped with such a special device, and a meeting does not always take place in a meeting room. It would be very useful if we could leverage the microphones built in the laptops some meeting participants bring for the meeting. Laptops are usually WiFi-enabled, so they can form an ad hoc network. Compared to traditional microphone array devices, such ad hoc microphone arrays are spatially distributed and the microphones in general are closer to the meetings participants. Thus, higher audio quality can be expected, assuming the microphones used in the laptops and those in the array devices have the same quality. On the other hand, ad hoc microphones present many challenges:

- Microphones are not synchronized;
- The location of the microphones/laptops is unknown;
- Microphones have different and unknown gains on different laptops; and
- The microphone quality is different, i.e., they have different signal to noise ratios.

Lienhart et. al. [4] developed a system to synchronize the audio signals by having the 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 by having each loudspeaker to play a coded chirp. Once the microphones are time synchronized and their positions are calibrated, traditional beamforming and sound source localization techniques can be used for speech enhancement, directing camera to speakers, and so on, to improve teleconferencing experience.

In this paper, we present an energy-based technique for locating human speakers (talkers). Compared to the previous technique by Raykar et. al. [5], our technique does not require accurate time synchronization, and it does not require the loudspeakers to send special coded chirps. In fact, we only use the average energy of the meeting participants' speech signals over a relative large window, so synchronization at 50 ms or even 100 ms suffices with our technique. The price to pay is that we cannot obtain position estimation as accurate as what was reported in Raykar et. al [5]. However, the position estimation is still good enough for many scenarios such as audio-visual speaker window selection in video conferencing [2, 6].

Given that the microphones are spatially distributed, a speaker/talker is usually relatively close to one of the microphones. Therefore, a simple mechanism for speech enhancement is to select the signal from the microphone that is closest to the speaker (or select the signal that has the best signal to noise ratio (SNR)). One problem is that the microphones have different gains thus resulting in abrupt gain changes in the output signal. We present an algorithm to estimate the relative gains of the microphones using meeting participants' speech signals. One nice property of the algorithm is that it is independent of the volume of the speakers so that each speaker can speak with any loudness he/she likes.

2. ENERGY-BASED SOUND SOURCE LOCALIZATION

Throughout this paper, we limit our discussions to the meeting room scenario where a number of meeting participants have their laptops in front of them. We assume that each laptop has a microphone and the laptops are connected by a network.

We first consider the case where every meeting participant has a laptop. We assume there are m laptops. For easy description, we assume each person speaks once. Let $y_i(t), i = 1, ..., m$ denote the audio stream captured by the i'th laptop. Let a_{ij} denote the average energy of the audio segment in $y_i(t)$ 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. Let c_{ij} denote the attenuation of person j's speech when it reaches laptop i. Let m_i denote the gain of the microphone on laptop i. We model a_{ij} as

$$a_{ij} = m_i s_j c_{ij} \tag{1}$$

We make the assumption that each speaker and its laptop are at the same location. Thus $c_{ij} = c_{ji}$, and $c_{ii} = 1$.

From equation 1, we have

$$\frac{a_{ij}}{a_{ii}} = \frac{m_i s_j c_{ij}}{m_i s_i} = \frac{s_j c_{ij}}{s_i} \tag{2}$$

and

$$\frac{a_{jj}}{a_{ji}} = \frac{m_j s_j}{m_j s_i c_{ji}} = \frac{s_j}{s_i c_{ji}} \tag{3}$$

Multiplying equations 2 and 3, we have

$$\sqrt{\frac{a_{ij}a_{jj}}{a_{ii}a_{ji}}} = \frac{s_j}{s_i} \tag{4}$$

Substituting equation 4 into 2, we have

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

Notice that equation 5 has the following properties: (1) it is independent of the laptop's gains, and (2) it is invariant of the scaling of the speech energy. For example, if a_{ji} and a_{ii} are multiplied by the same value, the right hand side remains the same.

Let d_{ij} denote the Euclidean distance between laptop i and j. Clearly c_{ij} is a function of d_{ij}. Theoretically speaking, audio energy is inversely proportional to the square of the distance between the sound source and the microphone. But our data indicates that d_{ij} is approximately a linear function of $\frac{1}{c_{ij}}$. Figure 1 is a plot of the relationship between $\frac{1}{c_{ij}}$ and d_{ij} based on the data recorded by having seven people, each with a laptop, to sit around a meeting table where end person was asked to speak a short sentence. d_{ij} are distances manually measured between the laptops by using a ruler. In total there are 21 data points and a Gaussian filter is used to smooth the curve slightly. We believe the reason we see a linear relationship is because of room reverberation, environmental and sensor noises, occlusions, and relatively small distances between the microphones and speakers.

Based on our practical observation, we set $d_{ij} = \frac{1}{c_{ij}}$ thus obtaining the distance between each pair of microphones.

We then use a technique of metric Multidimensional Scaling (MDS) [7] to obtain the 2D coordinates for each microphone.

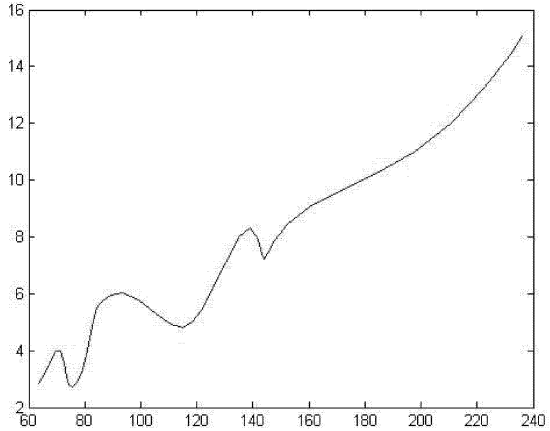

Fig. 1. The plot of $\frac{1}{c_{ij}}$ vs. d_{ij}

2.1. Speakers Without Laptops

In this section, we consider the situation where there are meeting participants who do not have laptops. Let us assume speaker k has no laptop. Note that we cannot apply equation 5 to compute c_{ik} anymore because a_{ki} and a_{kk} are not available. But we show that for any given two laptops i and j, we can compute the ratio $\frac{c_{ik}}{c_{jk}}$.

From equation 1, we have

$$\frac{a_{ik}}{a_{jk}} = \frac{m_i s_k c_{ik}}{m_j s_k c_{jk}} = \frac{m_i c_{ik}}{m_j c_{jk}} \tag{6}$$

Thus

$$\frac{c_{ik}}{c_{jk}} = \frac{a_{ik}}{a_{jk}}\frac{m_j}{m_i} \tag{7}$$

Again from equation 1, we have

$$\frac{a_{ji}}{a_{ii}} = \frac{m_j s_i c_{ji}}{m_i s_i} = \frac{m_j c_{ji}}{m_i} \tag{8}$$

Therefore

$$\frac{m_j}{m_i} = \frac{a_{ji}}{a_{ii}}\frac{1}{c_{ji}} \tag{9}$$

Substituting equation 9 into 7, we have

$$\frac{c_{ik}}{c_{jk}} = \frac{a_{ik}}{a_{jk}}\frac{a_{ji}}{a_{ii}}\frac{1}{c_{ji}} \tag{10}$$

Notice that c_{ji} can be computed from equation 5. Thus we are able to compute $\frac{c_{ik}}{c_{jk}}$ by using equation 10. Therefore the distance ratio is obtained by

$$\frac{d_{jk}}{d_{ik}} = \frac{a_{ik}}{a_{jk}}\frac{a_{ji}}{a_{ii}}\frac{1}{c_{ji}} \tag{11}$$

Let P_i and P_j denote the coordinates of laptop i and j, respectively. Notice that P_i and P_j can be computed by using the method described in the previous section. Let P_k denote the unknown coordinate of speaker k. Then we have

$$\frac{\sqrt{|P_k - P_j|^2}}{\sqrt{|P_k - P_i|^2}} = \frac{d_{jk}}{d_{ik}} \tag{12}$$

If there are m laptops. There are $\binom{m}{2}$ equations. When $m \geq 3$, we have enough equations to solve for the two coordinates of speaker k.

In our implementation, the system of equations 12 are solved by a nonlinear least square solver.

3. GAIN NORMALIZATION

In fact, Equation 9 is a formula to compute the gain ratios between any two microphones. To normalize the gains across the microphones, we just need to pick one of the microphones, say, microphone 1, as the reference microphone, and multiply the audio signal of the j'th microphone by $\sqrt{\frac{m_1}{m_j}}$.

4. EXPERIMENT RESULTS

We had 7 people each with a laptop sitting around a meeting table. The seven laptops have different brands. The laptop gains are set by the individuals arbitrarily. Each person was asked to speak a short sentence. Figure 2 shows the audio samples recorded from two of the laptops.

The seven audio files are roughly aligned by detecting the first speech frame through simple thresholding. 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 contribution of this paper.

The obtain the ground truth, we used a ruler to measure the distances between the laptops. For each laptop, we manually locate where the microphone is by visual inspection. We'd like to point out that the visual inspection may contain errors because it is difficult to determine where the microphone is for some laptops, and some laptops have multiple microphones. After we locate the microphone position for each laptop, we measure the distance between each pair of microphone locations. We then use the Multidimensional Scaling (MDS) [7] technique to compute the 2D coordinates from the measured distances. The coordinates are used as ground truth to evaluate the performance of our algorithm.

In Figure 3, the points marked as cross signs are the ground truth of the seven microphone positions. The points marked as circles are results estimated from our energy-based technique. The average error between the estimated positions and the ground truth positions is 22 centimeters.

We then simulate the situation where there are people who do not have laptops. In Figure 4, there are four laptops marked with circles. The rest of the three points (marked with cross signs) are speakers without laptops. The three plus signs are the positions estimated by our technique. Notice that only the 4 audio files recorded by the 4 microphones marked with circles are used to estimate the positions of the three speakers. We can see that the topology (who is close to whom) of the unknown speakers is

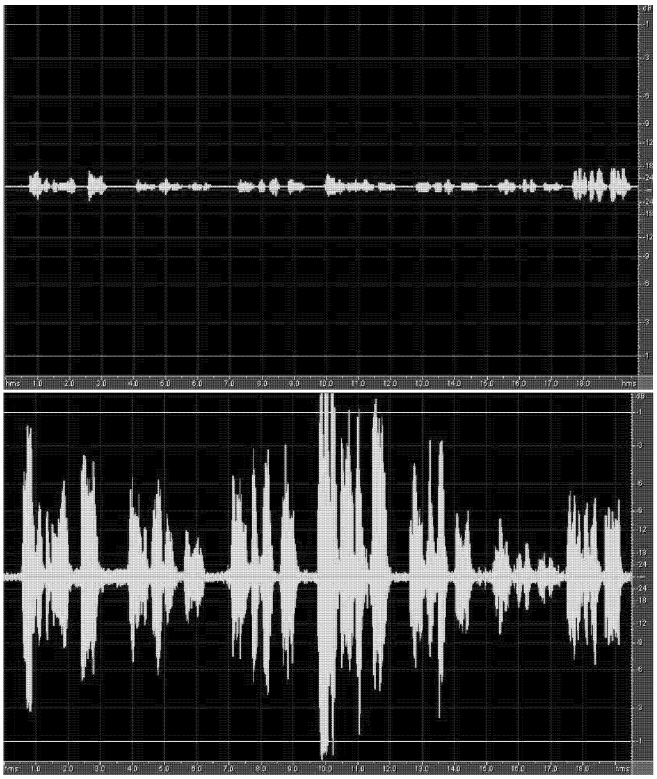

Fig. 2. Audio samples recorded from two of the laptops

estimated correctly. The estimated positions can be used for audio-visual speaker window selection during video conferencing since such systems do not require very accurate sound source localization [2, 6].

We have also experimented with gain normalization. For each speaker's audio segment, we select the microphone that has the best SNR and generate an aggregated audio stream. If we do not perform gain normalization, Figure 5 is the result. We can see that the audio energy contains many abrupt energy changes. In comparison, Figure 6 is the result after gain normalization. The audio energy level is consistent, and the difference between the energy levels of different speakers' speech segments in fact reflect the volume differences among the speakers.

5. CONCLUSIONS AND FUTURE WORK

We have presented an energy based technique to estimate the positions of the speakers from an ad hoc network of microphones. Our algorithm does not require accurate synchronization of the microphones, and it does not need to use special audio signals as in previously reported systems based on time of flight. This work is not intended as a replacement of the time of flight based approaches. In the situations where accurate time synchronization is difficult or it is not desirable to have laptops to emit special audio signals, our technique becomes a valuable alterative.

In addition, we presented a technique to normalize the gains of the microphones based on people's speeches. This is crucial to aggregate various audio channels from the ad hoc microphone

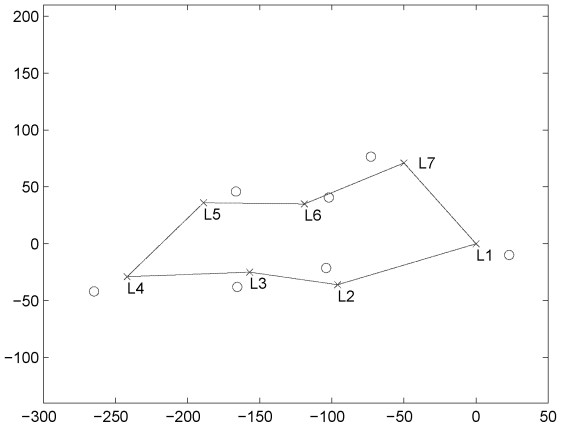

Fig. 3. The estimated positions of the seven laptops are marked in circles. The points marked with cross signs are positions estimated from the measured distances which are used as the ground truth.

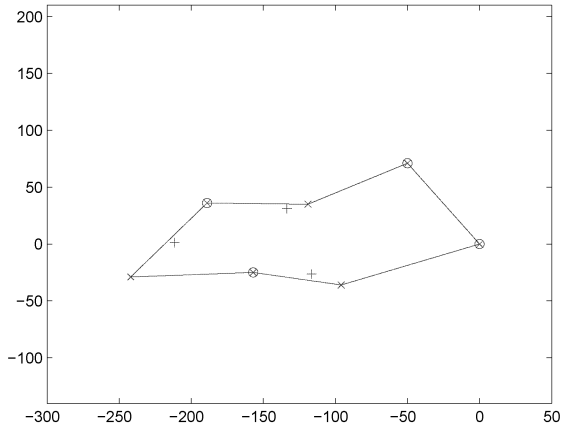

Fig. 4. Estimated positions of the three speakers who are assumed to have no laptops.

network into a single stream for audio conferencing. the technique has the nice property that it is invariant of the speaker's volumes thus making the system easy to deploy in practice.

6. REFERENCES

[1] "$http://www.polycom.com/products_services$," .

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

[3] M. Brandstein and H. Silverman, "A practical methodology for speech localization with microphone arrays," in *," Computer, Speech, and Language, 11(2):91-126, April*, 1997.

[4] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, "On the importance of exact synchronization for distributed audio processing," in *IEEE ICASSP 2003*, 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] C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola, "Boosting-based multimodal speaker detection for distributed meetings," in *IEEE International Workshop on Multimedia Singal Processing (MMSP) 2006*, 2006.

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

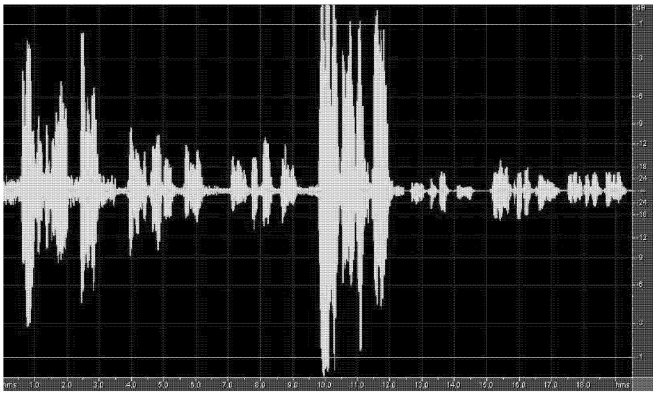

Fig. 5. Combined audio without gain normalization. The audio is combined by selecting the best-SNR segments among the seven microphones.

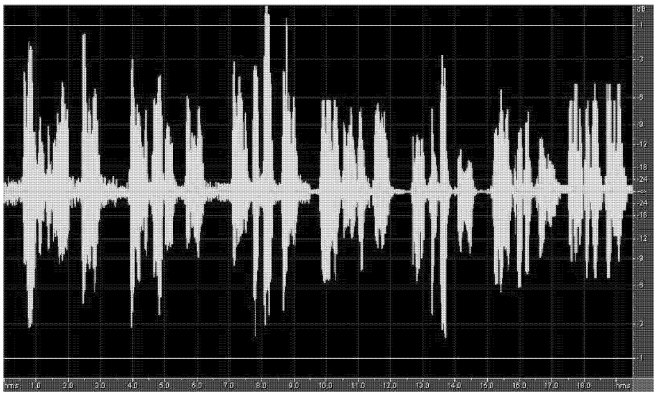

Fig. 6. Combined audio with gain normalization.