A BASELINE ALGORITHM FOR ESTIMATING TALKER ORIENTATION USING ACOUSTICAL DATA FROM A LARGE-APERTURE MICROPHONE ARRAY

J. M. Sachar and H. F. Silverman

LEMS, Division of Engineering, Box D. Brown University, Providence, Rhode Island 02912

ABSTRACT

We have shown that knowing the orientation of a talker in a large-aperture microphone array system can significantly improve location-estimation and beamforming algorithms. Measurements of a talker in an anechoic chamber have shown significant anisotropy in radiation patterns that may be used to influence the selection, processing, and weighting of microphone signals in such algorithms. Here we introduce a simple method for determining the orientation of a talker within a large focal area using only acoustic energy data obtained from the array. The mathematical basis for this procedure is presented and computed performance, based solely on acoustical measurements in a real environment, are listed and discussed.

1. INTRODUCTION

Large-aperture microphone arrays may be used to control audio for an audio teleconference and both audio and video for a video conference. Both applications require that the location and the orientation of a talker be known. To obtain clean audio signals, delay-and-sum beamformers require the talker's location[1, 2] or, equivalently, a delay parameter for each microphone. A talker's location is important for video conferencing so that a camera can be steered to the current talker.

The orientation of a talker is important when selecting among multiple cameras to ensure focus on the front and not on the back of a talker. It is also useful when beamforming to obtain quality audio. The convenient assumption that a talker is a point source is known to be suboptimal, in part, because there is added delay as the signal goes around the talker's head (head shadow). Therefore, if the orientation of the talker is known, a more sophisticated and accurate beamformer can be applied; for example, appropriately modifying the delays of the microphone signals having head shadow.

Past work has estimated the orientation/location (pose) of a talker by either using video alone [3, 4, 5] or using mixed video-acoustic approaches [6]. The problem with these methods is they require facial and head features to be available for evaluation and often fail if a talker's hair style is different from the model used, or if the contrast of the talker's skin relative to the background is inconsistent. Also, when a system exclusively uses audio signals, implementing visual algorithms would require adding a camera system.

The best microphone arrangement for using audio to determine the orientation of talkers over a large focal area is a large-aperture microphone array which completely surrounds talkers in the azimuthal plane. We shall refer to this type of array as having a surrounding aperture. The Huge Microphone Array (HMA) is capable of both recording and processing 512 microphones in real-time[7]. The current configuration of the HMA has 448 microphones distributed completely around a $4.5mx6.5mx3m$ laboratory environment forming a surrounding aperture in a plane parallel to the floor.

A human talker has a radiation pattern whose magnitude is not constant around the talker's head in either azimuth or elevation[8]. The pattern has been measured in anechoic chambers using both a torso simulator[8] and a set of human talkers[9]. The A-weighted pattern for speech, taken from [9], is shown for the azimuth angle in Figure 1. For our purposes, the elevation angle is a secondary effect and will not be considered here. It is clear from the figure and [9] that there is about a 6.6dB front-to-back ratio for the A-weighted signal for the average talker at most volume levels. This implies that one should be able to detect these differences in the energy pattern using the large-aperture array and thus obtain an accurate estimate of the talker's orientation. We call this the *energy method*. In particular, here, we try to determine the azimuth angle, θ.

There are also delay effects from the talker's radiation pattern that might be used for determining orientation[10]. However, the small changes to the times-of-arrival due to going around the head are difficult to determine from the microphone signals alone. A very precise knowledge of the true position of the talker's mouth is essential for success. In addition, the computational cost to obtain the delay-difference measurements is high relative to the cost of using a simple power average over a window of time. As determining talker orientation from a surrounding array is a relatively new topic, in this paper we introduce the details of the

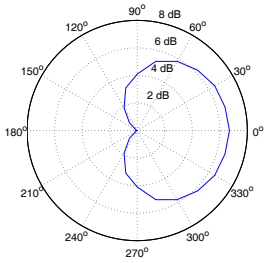

Fig. 1. A-weighted pattern for speech, taken from [9] at 0^o elevation

practical energy method to be used as a baseline for follow-on research and we show its performance experimentally using data from the real acoustical data. In this paper we will assume that a talker's location is known; only the orientation problem is examined. It should be noted, however, that the orientation and source location problem are intertwined.

2. A BASELINE APPROACH TO DETERMINING ORIENTATION: THE ENERGY METHOD

The energy method is probably the simplest and most straightforward way to determine a talker's azimuth angle. However, to exploit the radiation-pattern energy differences from data in a real environment, the method has to address the following four requirements:

1. Compensation for the inverse-square-law attenuation of the source signal
2. Reduction of the background noise effects
3. Enhancement of directional (higher frequency) components of the signal
4. Reduction of masking effects due to reverberations

Consider a simple, realistic model for the i^{th} microphone signal $m_i(t)$ for a single talker oriented at angle θ in azimuth $s(t,\theta)$ in a reverberant room with some uncorrelated background noise $n_i(t)$.

$$m_i(t) = h\left(t, \vec{x_i}, \vec{x_s}\right) * s(t,\theta) + n_i(t). \quad (1)$$

Here, $*$ indicates convolution and $h\left(t, \vec{x_i}, \vec{x_s}\right)$is the room impulse response from the source signal to microphone i. The vectors $\vec{x}_s$, the source position, and $\vec{x}_i$, the position of microphone i, are three-dimensional spatial vectors. If we decompose the impulse response into its direct and reverberant parts,

$$h\left(t, \vec{x_i}, \vec{x_s}\right) \equiv \frac{\delta(t-\tau_{is})}{d_{si}} + h_r\left(t, \vec{x_i}, \vec{x_s}\right), \quad (2)$$

where $d_{si} \equiv |\vec{x_s} - \vec{x_i}|$ then,

$$m_i(t) = \frac{s(t-\tau_{is},\theta)}{d_{si}} + h_r\left(t, \vec{x_i}, \vec{x_s}\right) * s(t,\theta) + n_i(t). \quad (3)$$

Our goal is to extract the directional source-signal energy. We first apply compensation for the inverse-square-law attenuation and time shift the microphone signal to obtain

$$\begin{aligned} \hat{m}_i(t) &\equiv m_i(t+\tau_{is}) \cdot d_{si} \\ &= s(t,\theta) + h_r\left(t+\tau_{is}, \vec{x_i}, \vec{x_s}\right) * s(t+\tau_{is},\theta) \cdot d_{si} \\ &\quad + n_i(t+\tau_{is}) \cdot d_{si}. \end{aligned} \quad (4)$$

If the second term (the reverberation noise) and/or the third term (background noise) is comparable to $s(t,\theta)$, then the desired dependence on angle might be masked. In addition, if the bulk of the energy in the signal is of low frequency, and thus non-directional, then we may not be able to find an orientation. However, we have seen that an A-weighted speech spectrum exhibits a useful directional pattern, implying using a highpass filter of this shape will address the latter problem. At the same time, most background noise (fan noise etc) is strongest at low frequencies. In Figure 2 we show spectra of the background noise in our environment, the A-weighting, and the designed filter, $f(t)$. We considered the trade-offs among maintaining A-weighting, eliminating noise, and preserving the speech content to design highpass filter $f(t)$. Applying $f(t)$ to the signal yields, ideally,

$$\begin{aligned} \bar{m}_i(t) &\equiv f(t) * \hat{m}_i(t) \\ &\approx \bar{s}(t,\theta) + h_r\left(t+\tau_{is}, \vec{x_i}, \vec{x_s}\right) * \bar{s}(t+\tau_{is},\theta) \cdot d_{si} \end{aligned} \quad (5)$$

where $\bar{s}(t,\theta)$ is the highpass-filtered version of $s(t,\theta)$.

We next compute the energy over a T-second window. Using the definition $E^T(xy) \equiv \int_0^T x(t)y(t)dt$.

$$\begin{aligned} E^T(\bar{m}_i^2) = E^T(\bar{s}^2(t,\theta)) &+ 2d_{si}E^T(h_r\left(t+\tau_{is}, \vec{x_i}, \vec{x_s}\right) \\ &* \bar{s}(t+\tau_{is},\theta) \cdot \bar{s}(t,\theta)) \\ &+ d_{si}^2 E^T([h_r\left(t+\tau_{is}, \vec{x_i}, \vec{x_s}\right) * \bar{s}(t+\tau_{is},\theta)]^2) \end{aligned} \quad (6)$$

or defining a few terms,

$$E^T(\bar{m}_i^2) \equiv \bar{S}(T,\theta) + R_i^1(T,\theta) + R_i^2(T,\theta).$$

If the energy method is to work, each of the last two terms has to be approximately constant in θ, significantly smaller in magnitude than the pattern differences of $\bar{S}(T,\theta)$, or vary in exactly the same fashion as $\bar{S}(T,\theta)$. We hypothesize that $R_i^1(T,\theta)$ will vary about zero rapidly as a function of microphone i and thus lowpass filtering $E^T(\bar{m}_i^2)$ in i will effectively remove its effects. $R_i^2(T,\theta)$ contains reverberation energy only. Reflections should be substantially attenuated due to a longer path length. Perhaps what is more

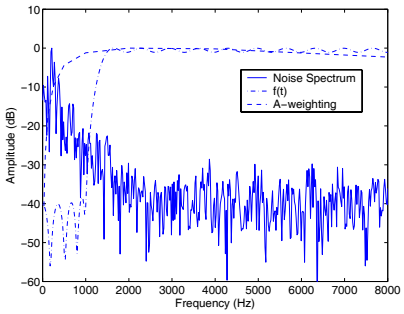

Fig. 2. Comparison of noise spectrum, A-weighting and filter $f(t)$

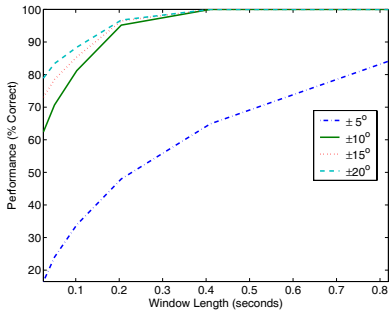

Fig. 3. Orientation Performance as a Function of Frame Size T for Four Allowed Orientation Tolerance Levels

significant is that $R_i^2(T,\theta)$ should be nearly constant as a function of θ as it is dependent a very large number of essentially random reflections from all directions.

The smoothing needed to remove $R_i^1(T,\theta)$ and preserve the desired shape from $E^T(\bar{m}_i^2)$ is not conventional because the azimuth angles from the source to the microphones are not in uniform increments with i. The microphones for the HMA system are placed randomly on vertices of a 3cm grid on two-dimensional panels mounted on walls; there are many cases in which several microphones are placed in a column. However, as we are lowpass filtering the data severely, these high frequency effects do not impact the result. We thus apply filter $l(i)$, a 127-point FIR filter having a normalized passband width of 0.005 which was down 55dB by 0.03. Finally, we select the azimuth orientation O as

$$O \equiv \underset{\theta}{argmax}\ [l(i) * E^T(\bar{m}_i^2)].$$

3. EXPERIMENTS/RESULTS

The method was tested using recordings of 4 seconds of speech from a human talker in a static position and facing a known direction. The HMA was used to record 448 simultaneously sampled microphones (and a close-talking microphone) distributed around the environment. In the first experiment, we wanted to see the performance relative to the frame size T, see Figure 3. The figure was obtained by running T-length windows over the entire length of the recording using a $10ms$ advance for $\pm 5^o, \pm 10^o, \pm 15^o$,and $\pm 20^o$ tolerances. It is clear that longer frame lengths are better, but, in a practical algorithm, length has to traded-off for algorithmic latency and/or computational practicality. If we want to be within $\pm 5^o$ and correct 60% of the time, then about a 400ms frame is needed. This size frame can also yield 95% correct results relative to $\pm 10^o$ tolerance.

Figure 4 is a polar plot of $E^T(\bar{m}_i^2)$ for a typical high-quality frame. The phenomena predicted in the last section are clear including the high-frequency noise as well as the apparent null to the back of the aimed direction. We also see

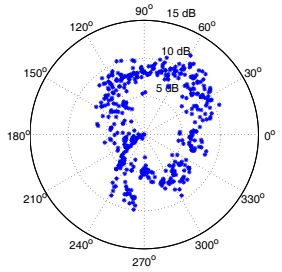

Fig. 4. Energy Radiation Pattern before Application of $l(i)$

some residual energy due to a large noise source(the HMA fans) at about 240^o. Figure 5 shows a top view of the environment with an overlaid polar plot of the same data after $l(i)$ had been applied. Note that the energy null is quite clear and the figure is quite similar to that predicted earlier, although made asymmetric by the residual energy from the noise source. The '·'s around the perimeter of the room represent microphones while the '*'s represent energy estimates as a function of θ.

To illustrate the kinds of speech for which the algorithm performs best, we used $T = 25.6ms$ and an error range of $\pm 20^o$ to show the 80% of the speech that performs the best. From Figure 6 it is pretty clear that having significant high-frequency energy in the source signal is the dominant factor. Note the spectrogram is for the close-talking microphone and has not been filtered in any way.

Finally, in a practical system, the location will likely be known with some error. Using $T = 0.8s$, some level of error was introduced on the surface of a sphere around the correct location and performance was measured. From Figure 7 the degradation in performance is shown. Performance for the largest tolerances did not change much when the location estimate was within 20cm of correct, but started deteriorating badly when the error was over 40cm. For $\pm 5^o$ tolerance, performance deteriorates about at the rate of about 1% per cm.

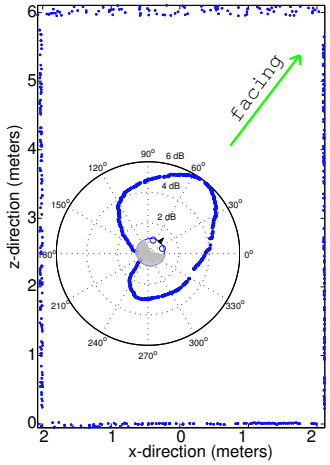

Fig. 5. Energy Radiation Pattern in a Top-Room Environment after Filtering by $l(i)$

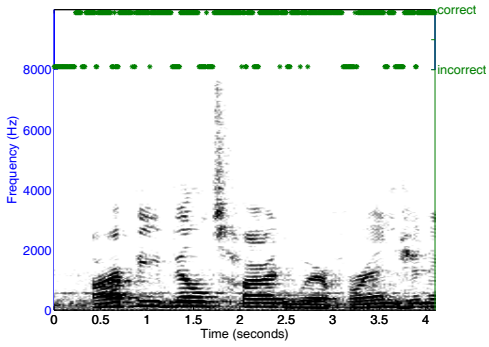

Fig. 6. Performance as a Function of Time for a 25.6ms Frame with $\pm 20^o$ Tolerance

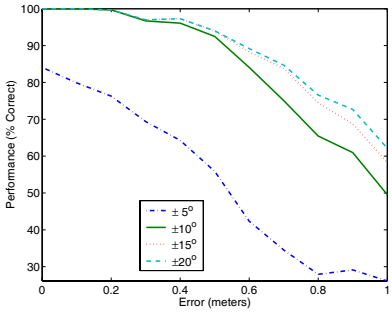

Fig. 7. Orientation Performance as a Function of Error in the Point-Source Location

4. CONCLUSION

We have developed a practical algorithm for determining human source orientation from large-aperture microphone array data. It is a baseline system and we expect further research using more sophisticated methods to offer improved performance, albeit at some increased computational cost. We observe 60% correct performance at the $\pm 5^o$ tolerance level or nearly 100% correct performance at $\pm 10^o$tolerance level using a 400ms time window. If estimates are made, say, every 50ms, this performance is adequate for a real-time system if we use some time averaging. The cost of the algorithm is quite small and is directly proportional to the number of points in the frame advance, if the implementation is done meticulously. Future work includes a real-time implementation, integration of this with a location estimation algorithm, and expansion of the method to include other orientation clues such as the extended time delays from those signals going around the head.

5. REFERENCES

[1] J. L. Flanagan, J. D. Johnson, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Amer.*, vol. 78, no. 5, pp. 1508–1518, November 1985.

[2] H. F. Silverman, W. R. Patterson III, J. L. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proceedings of ICASSP-1997*, Munich,Germany, April 1997, pp. I–251, I–254.

[3] R. Lopez and T. S. Huang, "Head pose computation for very low bit-rate video coding," in *6th International Conference on Computer Analysis of Images and Patterns*, Springer-Verlag Berlin Heidelberg, 1995, pp. 440–447.

[4] N. Kruger, M. Potzsch, and C. Malsburg, "Determination of face positions and pose with a learned representation based on labeled graphs," *Image and Vision Computing*, vol. 15, no. 8, pp. 665–673, August 1997.

[5] I. Shimizu, Z. Zhang, S. Akamatsu, and K Deguchi, "Head pose determination from one image using a generic model," in *3rd IEEE International Conference on Automatic Face and Gesture Recognition*, Nara Japan, April 1998, pp. 100–105.

[6] C. Wang and M. S. Brandstein, "Hybrid real-time face tracking system," in *ICASSP'98*, Seattle, Washington, USA, May 12-15 1998, vol. 6, pp. 3737–3741.

[7] H. F. Silverman, W. R. Patterson III, and J. L. Flanagan, "The huge microphone array (HMA)- Part I," *IEEE Transactions on Concurrency*, vol. 6, no. 4, pp. 36–46, October-December 1998.

[8] James L. Flanagan, "Analog measurements of sound radiation from the mouth," *The Journal of the Acoustical Society of America*, vol. 32, no. 12, pp. 1613–1620, December 1960.

[9] W.T. Chu and A.C.C. Warnock, "Detailed directivity of sound fields around human talkers," Tech. Rep. RR-104, National Research Council Canada, December 2002.

[10] H. F. Silverman, W. R. Patterson III, and J. M. Sachar, "Factors affecting the performance of large-aperture microphone arrays," *(In submission) Journal of the Acoustical Society of America*, 2001.