

# A NEW ALGORITHM FOR THE ESTIMATION OF TALKER AZIMUTHAL ORIENTATION USING A LARGE APERTURE MICROPHONE ARRAY

Avram Levi, Harvey F. Silverman

LEMS

Division of Engineering, Brown University Box D Providence, RI 02912

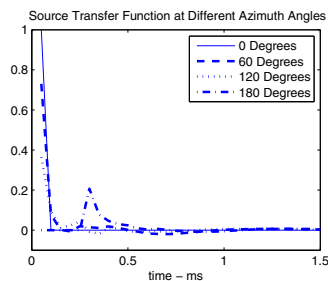
## ABSTRACT

Knowing the orientation of a talker allows a large-aperture microphone array to select and control cameras better in a teleconferencing situation, improve source-location estimation, and, often, improve beamforming. In 2004, we introduced a *baseline algorithm* for determining orientation azimuth. Recent testing showed the baseline algorithm behaved poorly when the source was not in the center of the focal area for the array. Here, we describe a second-generation algorithm, *A2*, that has overcome many of the *baseline's* shortfalls. It still extracts the estimate from microphone energies, but is improved by 1) using a narrow-band, high-frequency analysis, rather than the broad band of the *baseline algorithm*, 2) using spectral subtraction for uncorrelated noise removal and 3) fitting the processed microphone energies to an ideal model for the direct-wave energy. Most important is that 3) incorporates inverse-square-law effects properly on the direct wave **only**, which was not the case in the *baseline*. Results from an advanced simulator are presented to illustrate the issues. Then, *A2* and *baseline algorithm* results are compared using about 60 direct recordings from a human talker in a typical and noisy environment using our 448-microphone array. These show that *A2* is a significant improvement.

**Index Terms**— microphone array, talker orientation, acoustic energy measurement, reverberation, position measurement

## 1. INTRODUCTION

Large-aperture microphone arrays may be used to control audio and video components of a teleconferencing or audio-acquisition system. Determining the location and orientation of sources is an important part of the algorithms needed for such systems[4, 10]. Most of the past work has estimated the location/orientation of a talker by using either video alone[7, 6] or mixed video-acoustic approaches[11]. These algorithms require significant computational cost and are very dependent on the facial features of the talker. Also, there are cases in which a system has no video at all. In this paper we are concerned with finding the azimuthal orientation of a single talker in the focal area of a large-aperture microphone array in a typical noisy environment. We assume that a reasonable point-source estimate for the source location is known *a*



**Fig. 1.** Source model impulse responses parameterized by their angle off the normal to the mouth taken from the LEMS Simulator

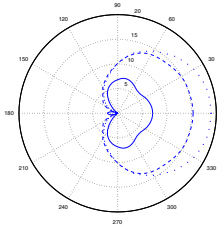
*priori*.

Using a linear-system assumption, we consider a discrete-time model for a signal received at the  $j^{\text{th}}$  microphone of a  $J$  microphone system,  $m^j(i)$ , at time-index  $i$  of a sampling interval of length  $T$  in a particular room.

$$m^j(i) = h_r^j(i) * h_s^j(i) * s(i) + n^j(i) \quad (1)$$

Here,  $*$  denotes convolution,  $s(i)$  is the talker's speech signal at the mouth,  $h_r^j(i)$  denotes the impulse response of the room from a fixed source to microphone  $j$ ,  $h_s^j(i)$  is the impulse response of the source itself for microphone  $j$  and  $n^j(i)$  is the uncorrelated background noise. The information for orientation in Equation 1 is all contained in the source model impulse response,  $h_s^j(i)$ . For a spherically radiating point source,  $h_s^j(i)$  is simply a unit impulse at  $i = 0$ . However, for a real human talker, while impulse-like, it has observable differences from an impulse. Using our simulator [1] that includes a head-shadow model and all the attenuation versus frequency data measured by Chu and Warnock [3] in an anechoic chamber, we show these effects in Figure 1. The task for finding the orientation is to isolate  $h_s^j(i)$  as closely as possible and then extract the orientation data from it. We shall concentrate on the azimuthal angle,  $\theta$ , only. Relative to a fixed coordinate in the room, we can note the dependence on azimuth angle as a function of the microphone index  $j$  as  $\theta_j$ . Thus the source impulse response may be denoted  $h_s^{\theta_j}(i)$ , and our task is to find

$$\theta_j^* \equiv \operatorname{argmax}_{1 \leq j \leq J} \Phi(h_s^{\theta_j}(i)) \quad (2)$$



**Fig. 2.** Example of a Measured Source Azimuth Radiation Pattern in an Anechoic Chamber at  $0^\circ$  Elevation at 1kHz(solid line), 3.15kHz(dashed line), and 8kHz(dotted line) [3]

where  $\Phi$  is some appropriate functional on the source impulse response.

One should note that the orientation of a talker affects the source impulse response through three mechanisms; 1) the energy radiation pattern as a function of frequency, 2) differences in time of arrival when the wave is slightly delayed by head shadow, and 3) effects of diffraction at the mouth. The latter two are relatively small effects and quite difficult to measure, especially in a noisy environment. However, for the energy pattern, very detailed measurements of the azimuth and elevation angle dependence as a function of frequency were carefully measured in an anechoic chamber by Chu and Warnock and presented in 2002[3]. Their measurements, and others, have shown that – see Figure 2 –, the energy is attenuated to the back of the talker[3, 5]. Thus, a talker is not a spherically-radiating point-source and the energy pattern might be sufficient to find azimuthal orientation.

In 2004, we reported a *baseline algorithm* [8] for determination of a talker’s azimuth angle when the source location is known. In it, as well as in this second-generation algorithm *A2*, we attempt to find the orientation from the source energy radiation pattern only. The problem breaks down into two separate parts. First, some processing needs to be done to isolate, as much as possible, the source impulse response or transfer function. In the second part, the approximated source transfer function is used to determine the azimuthal orientation.

## 2. PROCESSING TO ISOLATE THE SOURCE TRANSFER FUNCTION

Consider the model in Equation 1. The first step in obtaining an isolated representation of  $h_s^{\theta_j}(i)$  is to eliminate the uncorrelated background noise. In the *baseline* system, this was done by a fixed highpass filter. In *A2* we use standard spectral subtraction[2], assuming we can obtain a reasonable background noise estimate from a quiet portion of a test utterance. As spectral subtraction is done in the frequency domain, we shall use the discrete frequency-domain representation for samples indexed by  $r$  representing frequencies  $2\pi r/T$  where

$T$  is the sampling interval and  $r \in [0, N - 1]$  for the  $N$ -point DFT, and try to obtain an estimate of the source transfer function,  $H_s(r, \theta_j)$ . Of course, while spectral subtraction does a better job, in general, than did the fixed filter in the *baseline* system, the result is imperfect. Be that as it may, we opt to continue with the derivation considering that there is no longer any uncorrelated noise, i.e.,  $n_j(i) = 0$ . Breaking up the room transfer function into its direct-wave and reflected-wave subcomponents,  $Hr(r) = H_{dir}(r) + H_{rev}(r)$ , the discrete-time Fourier transform of each microphone signal becomes,

$$M_j(r) = H_s(r, \theta_j) \cdot (H_{dir}^j(r) + H_{rev}^j(r)) \cdot S(w) \quad (3)$$

As we know the point-source estimation for the source location, we can compute the appropriate time shifts to make certain that the direct signal received for a particular frame of data is for the same interval of speech for each microphone. Then we can determine the energy in a frame,  $l$ , for each microphone by taking the sum of the squares of the time samples within the frame. This yields

$$E_j(r) = |H_s(r, \theta_j)|^2 \cdot |S_l(r)|^2 \cdot \{ |H_{dir}^j(r)|^2 + |H_{rev}^j(r)|^2 + 2|H_{dir}^j(r)||H_{rev}^j(r)| \cos(\gamma(r, j)) \} \quad (4)$$

where  $\gamma(r, j)$  is the phase difference between the direct and the reverberant room responses. We define the terms as follows:

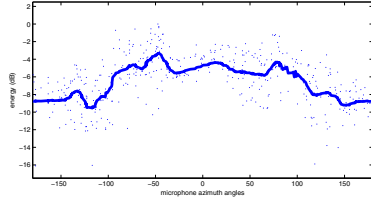
$$E_j(r) \equiv E^{dir}(r, \theta_j) + E^{rev}(r, \theta_j) + E^{mix}(r, \theta_j) \quad (5)$$

We now have  $E_j(r)$ , energy estimates for each frequency of each microphone for frame  $l$ . If the room were anechoic, then  $E_j(r) = E_j^{dir}(r, \theta_j)$  only and as the direct wave room impulse response is a constant  $B$  times a unit impulse at  $i = 0$ , (remember the data was time adjusted earlier for framing), making its DFT a constant then,

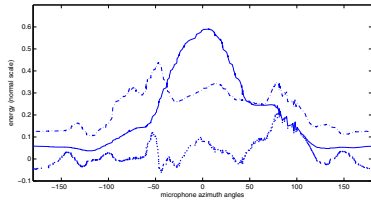
$$E_j^{anechoic}(r) = B|H_s(r, \theta_j)|^2|S_l(r)|^2. \quad (6)$$

In this ideal case, we would have our estimate of the magnitude of the transfer function of the source. We could use  $E_j(r)$  to find a maximum at one(or more) excited frequencies which should indicate the talker’s front, minimum, which should indicate a talker’s back, or even do some fitting that used all the information available.

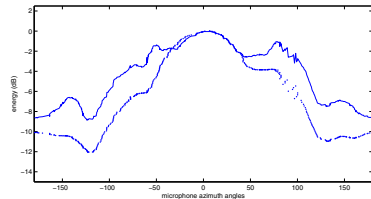
However, we have the other two terms of Equation 5 in the real case in which the room has reflections. All three terms making up  $E_j(r)$  multiply the source transfer function, and as a result have some potentially useful information for our purposes. What we needed to determine was a mechanism that would extract as much information from these two terms as possible that would distort our later search for the azimuthal direction from  $E_j(r)$ . In the *baseline* system, we found that



(a)  $E^{rev}(r\Delta f = 3150Hz, \theta_j)$  - before (dots) and after smoothing (solid line)



(b)  $E^{dir}(r\Delta f = 3150Hz, \theta_j)$  (solid line),  $E^{mix}(r\Delta f = 3150Hz, \theta_j)$  (dotted line) and  $E^{rev}(r\Delta f = 3150Hz, \theta_j)$  (dashed line) all after smoothing. Note: The mix term can be negative so the plot is on a relative power scale, instead of dB.



(c)  $E_j(r)\Delta f = 3150Hz$  (solid line) and  $E^{dir}(r\Delta f = 3150Hz, \theta_j)$  (dotted line) after filtering

**Fig. 3.** An Example of the Three Energy Components,  $E_j^{dir}(r, \theta_j) + E_j^{rev}(r, \theta_j) + E_j^{mix}(r, \theta_j)$ . Behavior of the energy terms vs azimuth angle at 3150Hz

a severe lowpass-filter smoothing with respect to the azimuth angle yielded a useful energy pattern. We only suggested the reasons why this was the case. Now, however, as we had the improved simulator, we could at least test the idea on suitable simulated data for a 200Hz wide band of frequencies about 3150Hz in which we assume the speech level to be constant over the frame. The results are shown in Figure 3. In Figure 3a)  $E^{rev}(r\Delta f = 3150Hz, \theta_j)$  is shown both before and after the filtering. There clearly is a large amount of "high-frequency" noise on this energy signal which is likely due to the standing-wave patterns of the largest reflections for this high-frequency narrow-band data. The filtered reverberant term has a broad peak around the right azimuth angle, and shows variation which is small when compared to the other terms. In Figure 3b), the three terms are shown on a normalized energy scale, not dB, because the mixed term can have negative values. One can see that the direct term dominates at the correct orientation value and that the mixed term is rel-

atively small – the components tend to cancel out. The reverberant term adds at the correct orientation, but introduces significant noise in the directions off of the peak. In Figure 3c) we compare the smoothed energy estimate to the actual direct term. One can notice the distortion of the signal due to the reverberant and mixed terms being added to the direct, but the correct orientation is clearly discernable. In A2, the filtering to, as much as possible, isolate the direct energy term was more or less the same as in the *baseline system*.

### 3. EXTRACTING THE ORIENTATION ESTIMATE FROM $E_j(r)$

A major error source in the *baseline system* was that, after smoothing, we corrected the energy estimates for inverse-square-law attenuations. As one can see from Equation 5, this procedure is incorrect for the mix and the reverberant terms. As a result, the *baseline system* often failed for this reason when the source was not in the center of the focal area of the room.

In A2, we construct an ideal model for the direct energy term by using the data from Warnock and Chu [3]. That is, for each potential azimuthal direction indexed by  $\theta_n$  using the known source point – either to each of the 448 microphones in our system or, after some interpolation, to evenly-spaced angles (about  $1^\circ$ ) – develop an idealized energy estimate  $E_{\theta_n}^{ideal}(r, j)$ . We obtain the power from the Chu and Warnock data for the given direction, divide by the distance, and normalize so that its maximum value is the same as the value of the direct energy estimate at that maximum value angle. Then we have  $E_{\theta_n}^{ideal}(r, j)$  and, for the noninterpolated situation, obtain the azimuthal estimate for some frequency (or frequency band)  $r$  from,

$$\theta_n^* \equiv \operatorname{argmin}_{1 \leq n \leq N} \sum_{j=1}^J (w(\theta_n, r, j)(E_{\theta_n}^{ideal}(r, j) - E_j(r))^2) \quad (7)$$

The interpolated case is similar. The weighting,  $w(\theta_n, r, j)$  needs to be nonuniform to compensate for the effects due to reverberation that were illustrated in Figure 3c. We tested a few and ended up using

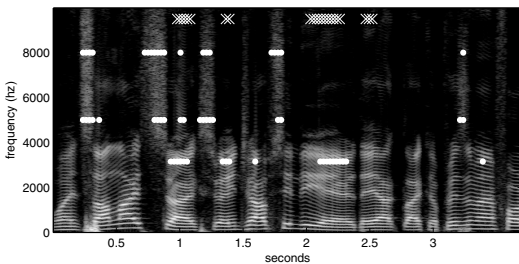
$$w(\theta_n, r, j) = E_{\theta_n}^{ideal}(r, j)^2. \quad (8)$$

### 4. EXPERIMENTS

The algorithm was evaluated by using a set of 4-second recordings of speech from a **real, human** talker at 25 positions in the room's focal area and three orientations for each position. Some of the combinations of position and orientation were pathological, e.g., talking right into a wall, so the number of datasets was reduced from 75 to 59 realistic combinations. Our 448-microphone, 20kHz sampling rate, Huge Microphone Array (HMA) [9], augmented with a

close-talking microphone channel, was used to record each 70MB dataset. (These are available for comparable research at [www.lems.brown.edu/array/datasets](http://www.lems.brown.edu/array/datasets)). One should note that this test is far more complete and difficult than the one used in [8]. Note also, that the beginning of each recording was a half-second of silence that could be used for the spectral-subtraction algorithm.

The work in [3] tabulates the magnitude of the source transfer function at 1/3-octave frequencies, ranging from 160Hz to 8000Hz. In A2 we evaluate in 200Hz wide, narrow bands. As high-frequency components are the most directional, we used 3150Hz, 5000Hz, and 8000Hz as center frequencies. Thus we wanted to make decisions only for those frames having some energy at those frequencies. Here, we avoided developing a discriminator, a necessary component of a real-time algorithm, and just selected the top 20 of 140 51.2ms overlapping frames for each frequency or the 20 frames having the highest broadband high-frequency energy for the *baseline*. The data are shown in Figure 4. One should note that the data around 8kHz is all from the fricated sounds *s,z*, and one fricated vowel. The data for 5kHz is about the same, except for the burst in a *d*. The data for 3.15kHz is mainly vocalic high formants. Cumulative results for the 59

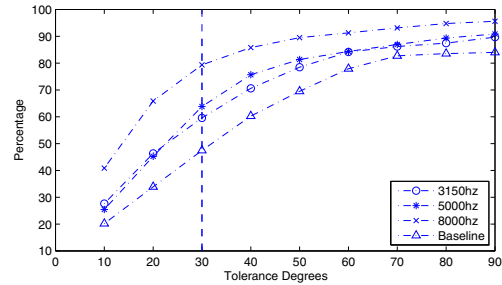


**Fig. 4.** Spectrogram of close talking microphone data along with the chosen frames (white dots=improved algorithm frequencies, white x's=baseline algorithm band). The speech uttered is "When sunlight strikes raindrops in the air, they act like a prism and fo..."

datasets are shown in Figure 5. Given that subsequent algorithms require a somewhat loose estimate of the azimuthal angle, we considered correctness to be within  $\pm 30^\circ$ . From Figure 5, we see that the 8kHz data are correct about 80% of the time and the 5kHz and 3.15kHz data are correct about 60% of the time. These results are significant improvements over the *baseline* algorithm. It was somewhat surprising that the best result was at the highest frequency where the amount of energy is smaller, but it seems that the effects of reflections and background noise are comparatively reduced at this frequency and the directivity of a human talker is narrower at 8kHz.

## 5. CONCLUSION AND FUTURE WORK

We have presented and compared an improved algorithm (relative to [8]) for determining azimuthal talker orientation. The



**Fig. 5.** Percentages of found angles within the specified tolerances - 200Hz Bandwidth

improved algorithm makes extensive use of the source characterization data described in [3] and features a narrow-band analysis. Given a  $30^\circ$  tolerance, we are able to determine correct azimuthal orientation about 80% of the time at our best frequency, about 35% better than in the *baseline* method over a pretty complete set of talker locations and orientations in a very noisy and reflective room. There is still plenty of work to do! We are currently working on creating a robust version of the algorithm using some local beamforming ideas.

## 6. REFERENCES

- [1] H. S. Avram Levi. Lems acoustic simulator. In <http://www.lems.brown.edu/array/download.html>, 2007.
- [2] S. F. Boll. Suppressing of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, April 1979.
- [3] W. T. Chu and A. C. C. Warnock. Detailed directivity of sound fields around human talkers. Technical Report IRC-RR-104, Institute for Research in Construction, National Research Council, Canada, 2002. <http://irc.nrc-cnrc.gc.ca/ircpubs>.
- [4] 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.*, 78(5):1508–1518, November 1985.
- [5] M. Kompis and N. Diller. Simulating transfer functions in a reverberant room including source directivity and head-shadow effects. *J. Acoust. Soc. Amer.*, (5):2779–2787, May 1993.
- [6] 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*, 15, 1997.
- [7] 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*, pages 440–447, Springer-Verlag Berlin Heidelberg, 1995.
- [8] J. M. Sachar and H. F. Silverman. A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array. In *ICASSP'04*, volume 4, pages 65–68, 2004.
- [9] H. F. Silverman, W. R. Patterson, and J. L. Flanagan. The huge microphone array (HMA)- Part I. *IEEE Transactions on Concurrency*, 6(4):36–46, October-December 1998.
- [10] H. F. Silverman, W. R. Patterson, 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*, pages 251–254, Munich, Germany, April 1997.
- [11] C. Wang and M. S. Brandstein. Hybrid real-time face tracking system. In *ICASSP'98*, volume 6, pages 3737–3741, Seattle, Washington, USA, May 1998.