IMPROVING AUDIO SOURCE LOCALIZATION BY LEARNING THE PRECEDENCE EFFECT

Kevin Wilson *Trevor Darrell*

Vision Interface Group, CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

Audio source localization in reverberant environments is difficult for automated microphone array systems. Certain features observable in the audio signal, such as sudden increases in audio energy, provide cues to indicate time-frequency regions that are particularly useful for audio localization, but previous approaches have not systematically exploited these cues. We learn a mapping from reverberated signal spectrograms to localization precision using ridge regression. The resulting mappings exhibit behavior consistent with the well-known precedence effect from psychoacoustic studies. Using the learned mappings, we demonstrate improved localization performance.

1. INTRODUCTION

Source localization is an important basic problem in audio processing, but existing algorithms perform poorly in reverberant environments [1]. Techniques that assume an anechoic environment become much less reliable in reverberant environments, and techniques that try to compensate for the reverberation by learning a dereverberating filter are very sensitive to even small changes in the acoustic environment [2].

To allow for source motion, most localization systems compute localization cues based on short time segments of a few tens of milliseconds and combine these cues using a source motion model. In such systems, either the low-level cues themselves can be improved, or the means by which the cues are combined can be improved. This paper focuses on the latter area, learning an improved uncertainty model for the low-level cues. We use cues from the reverberated audio to predict the uncertainty of localization cues derived from small time-frequency regions of the array input. Any localization cue can be used with our approach, but in this paper we use time delay of arrival (TDOA) estimates based on cross-correlation in a set of time-frequency regions.

We make three contributions. First, we devise a method that uses recorded speech and simulated reverberation to generate a corpus of reverberated speech and its associated localization error. Second, we use this corpus to learn mappings from the reverberated speech to a measure of localization uncertainty and demonstrate its utility in improving source localization. Third, we make a connection between the mappings learned by our system and the precedence effect, the tendency of human listeners to rely more on localization cues from the onsets of sounds. While other systems, such as [3] and [4] have employed heuristic mappings or mappings that approxmate the ML weighting, we believe that this paper is the first attempt to learn such a mapping from a training corpus.

2. BACKGROUND

2.1. Array processing

Cross-correlation is a standard technique for TDOA estimation in array processing. To estimate a TDOA between two microphones, the two signals are cross-correlated, and the lag corresponding to the maximum cross-correlation is assumed to be the TDOA. Attempts to improve TDOA performance in reverberant environments fall into two broad categories – some systems attempt to build in robustness to reverberation at a very low level and some attempt to improve the way in which multiple localization cues are fused into a final location estimate.

In the first category, [1] reviews much of the work relevant to microphone arrays. In particular, filtering the signals before cross-correlating can increase robustness to reverberation. The phase transform, in which the microphone signals are whitened before cross-correlation, is one popular technique for increasing robustness to reverberation. After whitening, no single frequency dominates, and that the effects of reverberation cancel out when averaged over many frequencies. Another technique is to use the ML solution for TDOA estimation by doing cross-correlation after applying a filter that weights each frequency according to its SNR. In practice, however, the SNR is usually unavailable. Another approach is to use detailed models of the reverberation to undo its effects. [5] learned detailed models of the cross-correlation waveforms corresponding to a small set of training locations in a room, but no results were presented to suggest how well the approach generalized to novel locations. [2] shows that the fine structure of the reverberation effects in a room can vary greatly and unpredictably even over distances of tens of centimeters, so it is unclear how robust methods in this thread can be.

In the second category, [4] trained a neural network to fuse multiple audio and visual cues to localize a sound source, and [3] engineered a number of heuristics, including a simple version of the precedence effect, into a system for combining multiple audio localization cues. These systems demonstrate the potential for improving cue fusion; however, [4] used only a few audio features to control fusion, and it is unclear how the heuristics in [3] were chosen. Our technique falls into this category, and it provides a principled way of fusing cues based on mappings learned from a training corpus.

2.2. The precedence effect

The precedence effect, also known as the "Haas effect" or the "law of the first wavefront," is the psychoacoustic effect in which the

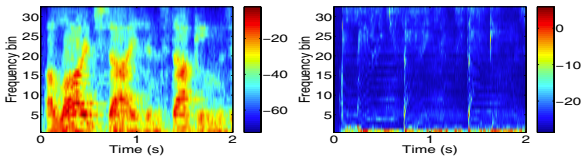


Fig. 1. Empirical justification for the precedence effect. On the left is a spectrogram of the reverberant speech received at one of the microphones in the array. On the right is the corresponding map of the empirical localization precision (in dB) for each time-frequency bin. Sudden onsets in the spectrogram, such as those at 0.7 seconds and 1.4 seconds, correspond to time-frequency regions with high localization precision.

apparent location of a sound is determined largely by the localization cues from the initial onset of the sound [6, 7]. It has been argued that the precedence effect improves people's ability to localize sounds in reverberant environments. Because direct path sound arrives before any reflections, initial onsets will tend to be less corrupted by reverberation than subsequent sounds.

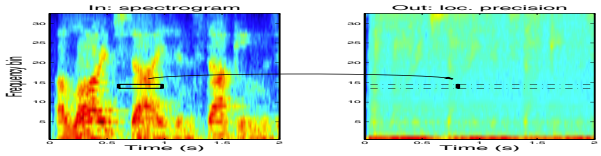
Although the basic purpose of the precedence effect seems straightforward, the details are not clear. The notion of an "onset" is imprecise, although recent progress has been made in [8] in determining the time-scales over which the precedence effect operates. In addition, most studies have focused on stimuli such as click trains or noise bursts, and it is unclear how to apply their findings to more natural sounds. Studies on human infants and young puppies (reviewed in [6]) found no evidence of the precedence effect, and studies on young children have found the effect to be much smaller. Together with the studies of adults, this suggests that the precedence effect may be learned over the first few years of life. The imprecision of the standard description of the effect and the possibility that children learn the precedence effect suggest that it may be fruitful to apply a learning approach to the problem of audio source localization in reverberant environments.

3. METHODS

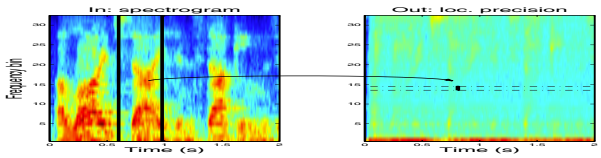
Our goal is to learn an association between the audio signal and the localization precision, which we define to be the reciprocal of the empirical localization error. To do so, we generate a training corpus consisting of a set of spectrograms of reverberated speech signals and a time-frequency map of the localization precision over the course of these speech signals as shown in Figure 1. We then compute a set of filters that estimate the localization precision from the spectrogram representation of the reverberated audio.

3.1. Corpus generation

We generate the training corpus by using the image method of reverberation modeling [9] to simulate a room containing one speech source and two microphones. The simulation, which treats each wall of the room as a sound "mirror" with a frequency-dependent absorption coefficient, includes the effects of reverberation, and we add stationary noise to model sounds such as computer fans and ventilation systems. We synthesize N_r realizations of the utterance, each with the speech source and microphones in a different location in the room, and calculate the localization precision over all realizations.



(a) Narrowband precision calculation



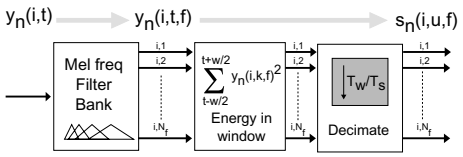
(b) Broadband precision calculation

Fig. 2. An illustration of the narrowband and broadband mappings for frequency band 15. In 2(a) an FIR filter estimates the localization precision as a function of spectrogram bin 15. In 2(b) an FIR filter estimates the localization precision as a function of all spectrogram bins. The dashed lines indicate that to estimate confidence for all times, the mapping is applied at all time offsets.

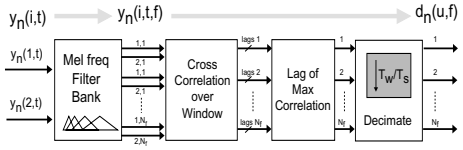
More formally, we start with a single speech signal, $x(t)$, and randomly generate N_r simulated room configurations. We represent these configurations as filters $H_n(i,t)$, where $n \in \{1 \dots N_r\}$ represents the room realization and $i \in \{1,2\}$ represents the i^{th} microphone signal. Passing $x(t)$ through $H_n(t,i)$ and adding noise signal $z_n(i,t)$ yields $y_n(i,t)$, a set of reverberated speech signals. We then pass $y_n(i,t)$ through a mel-scaled filter bank, yielding $y_n(i,t,f)$, $f \in \{1 \dots N_f\}$ where N_f is the number of bands in the filter bank. We calculate cross-correlations between time-windowed segments of the two channels of $y_n(i,t,f)$, and assign the time delay of the maximum cross-correlation to $d_n(u,f)$, where frame index u is replaces the time index t. Finally, we calculate $e(u,f) = \frac{1}{N_r}\sum_{n=1}^{N_r}(d_n(u,f) - d_{n_{true}}(u,f))^2$, the localization error variance, and $p(u,f) = -10 * log_{10}(e(u,f))$, the localization precision (in dB). We calculate a speech spectrogram, $s_n(i,u,f)$ from $y_n(i,t,f)$ by calculating the energy (in dB) of time-windowed segments. Figure 3 contains block diagrams describing these calculations.

3.2. Filter learning

We then use ridge regression [10] to learn FIR filters that estimate the localization precision (in dB) from the reverberated spectrogram (in dB). In this paper, we examine two different forms for these filters. In the first case, which we call a narrowband mapping, we learn a separate FIR filter from each frequency band in the spectrogram to the corresponding frequency band in the localization precision output as shown in Figure 2(a). In the second case, which we call a broadband mapping, we learn a separate FIR filter for each band of the localization precision output, but in each case the input comes from all frequencies of the input spectrogram. This case is illustrated in Figure 2(b). We chose to examine the narrowband case because, for the case of stationary signals, each frequency band is uncorrelated with all other frequency bands, and thus the narrowband mapping should be sufficient in this case. Al-



(a) Speech spectrogram calculation



(b) TDOA calculation

Fig. 3. Spectrogram and TDOA processing block diagrams (see text).

though speech is nonstationary, this narrowband mapping provides a useful baseline against which to compare. The broadband mapping subsumes the narrowband mapping and should be able to capture cross-frequency dependencies that may arise from the nonstationarity of speech.

For the narrowband mapping with causal length l_c and anticausal length l_{ac}, we solve N_f regularized linear least-squares problems of the form $\mathbf{z_f} = \mathbf{A_f}\mathbf{b_f}$, $f \in \{1 \dots N_f\}$ where

$$\mathbf{z_f} = (\dots p(u,f)p(u+1,f)\dots)^{\mathsf{T}}$$

$$\mathbf{A_f} = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ s(u-l_c,f) & s(u+1-l_c,f) & \dots & s(u+l_{ac},f) & 1 \\ s(u+1-l_c,f) & s(u+2-l_c,f) & \dots & s(u+1+l_{ac},f) & 1 \\ s(u+2-l_c,f) & s(u+3-l_c,f) & \dots & s(u+2+l_{ac},f) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad (1)$$

and $\mathbf{b_f}$ is an FIR filter with $(l_c + l_{ac} + 1)$ taps stacked with a DC component.

For the broadband mapping, we solve N_f regularized linear least-squares problems of the same form as in the narrowband case, but with each row of $\mathbf{A_f}$ consisting of a concatenation of spectrogram values from all frequency bands and with b_f representing an FIR filter with $(l_c + l_{ac} + 1) * N_f$ taps stacked with a DC component. For both types of mapping, the ridge regression λ parameter is set through cross validation.

When applying this technique to localization, the only computational costs (beyond the basic TDOA calculations) are of computing a spectrogram on the incoming audio signal and applying a set of short FIR filters to that spectrogram. Because the signals that we regress between, the spectrogram and the mean square error, do not depend on the detailed structure of the reverberation, our technique is robust to changes in location in the room.

4. RESULTS

In this evaluation, we use audio sampled at 8 kHz, and we use a mel-scaled filter bank with $N_f = 30$ frequency bands centered from 333 Hz to 3700 Hz. The frame rate for our spectrogram and for our TDOA estimates is 267 frames per second. We use 17 minutes of speech for training, and a separate 90 seconds of speech for testing. Our simulated room is roughly 4m $\times$ 7m $\times$ 2.3m and has a reverberation time of roughly 300 ms.

Method	RMS TDOA error (ms)	RMS angular error (degrees)
True precision	0.28	13.9
Broadband mapping	0.29	14.3
Narrowband mapping	0.31	15.4
Scalar mapping	0.31	15.4
Uniform weighting	0.35	17.4

Table 1. Root-mean-square (RMS) localization error for different learned mappings. The broadband filter achieves an error nearly as small as is achieved using the true (empirically determined) precision.

4.1. Localization results

Table 1 shows the decrease in localization error achieved by our technique. Test data, generated from different utterances and in a different location than any of the training data, was synthesized in same simulated room used for the training data, generating a test spectrogram, $s_{test}(u,f)$, and a set of test TDOAs, $d_{test}(u,f)$. The mappings learned according to the method in Section 3.2 were applied to $s_{test}(u,f)$, yielding an estimated localization precision map, $p_{est}(u,f)$. Assuming independent errors in different time-frequency regions, the minimum-variance estimate of the TDOA is to take a weighted mean of all $d_{test}(u,f)$, where the weights are proportional to $p(u,f)$. Since we do not have access to the true $p(u,f)$ in practice, we use our $p_{est}(u,f)$. Table 1 shows root-mean-square (RMS) localization error achieved by each method when fusing TDOA estimates over 0.5 second audio segments. The angular error associated with this TDOA error depends on the array geometry; numbers in the table assume a microphone spacing of 40 cm.

Each row shows the performance of a different method of estimating precision information. The first row, "True precision," shows results using the empirically determined (ground truth) precision of each time-frequency region in the test set. This is the best that can be done (under our independent error assumption) with the given localization cues and acoustic environment. "Broadband mapping" and "Narrowband mapping" are the mappings described above. "Scalar mapping" is a simple special case of the narrowband filter using only one tap. "Uniform weighting" uniformly weights each time-frequency region; it corresponds roughly to the phase transform described in Section 2. In all cases, variants of our technique outperform the uniform weighting, and the broadband mapping achieves nearly the same error as using the empirically determined precision.

All of the errors, including the error when using ground truth, are large because we are computing separate TDOA estimates for each frequency band and combining them using our estimated precisions. We chose to do this because it has a straightforward interpretation in which we are combining multiple estimates with known variances to compute an optimal estimate, but it does not achieve the best possible performance. We are currently working on using our estimated precisions in a generalized cross-correlation framework, in which our weightings do not have as obvious an interpretation, but which should yield lower errors for all experimental conditions.

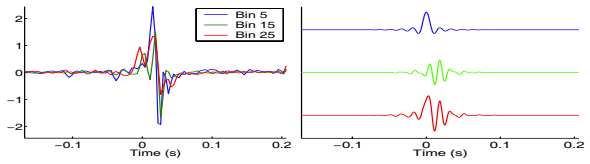


Fig. 4. Narrowband filters. Left shows a representative subset of the learned filters. Right shows a schematic decomposition of the learned filters. Each of the narrowband filters on the left can be viewed as a linear combination of a low-pass filtered impulse (top) with a band-pass filtered edge detector (middle). The bottom curve shows the linear combination of the top two curves, which is qualitatively similar to the filter for bin 25.

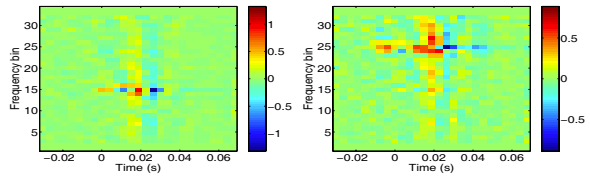


Fig. 5. Learned broad-band filters for two filter bands. These filters have most of their energy in the frequency bin whose precision they are estimating, but because of the non-stationarity of speech there is energy across many frequency bins. Left is frequency bin 15, centered at 1316 Hz. Right is frequency bin 25, centered at 2618 Hz.

4.2. Relationship to the precedence effect

Figure 4 (left) shows the learned FIR filters for a representative subset of the filter bands. In all three cases the filter is approximately a superposition of a low-passed delta function and a band-passed edge-detector, as depicted schematically in Figure 4 (right). The low-passed delta function component indicates that louder sounds provide better localization cues, which is to be expected in the presence of additive noise, where the ML frequency weighting is proportional to the SNR and the SNR in our scenario is roughly proportional to the signal energy. The band-limited edge-detector can be interpreted as an onset detector, which is consistent with the precedence effect that has been studied extensively in psychoacoustics. The relative amplitudes of the impulse and the edge detector reflect the relative importance of these two effects at each frequency. In our scenario, SNR effects dominate at low frequencies, while precedence-like effects dominate at higher frequencies.

Our results are qualitatively similar to the maximum likelihood frequency weighting and the precedence effect, but they go beyond that by learning structure that is specific to the speech signal itself. For example, while the broadband mappings are mostly localized around the frequency whose localization precision they are estimating, there is energy across the entire spectrum in some of the filters, most obviously in bin 25 in Figure 5 (right). Additionally, while there have been studies of the time-scales over which the precedence effect operates, most of these have used simple sounds such as click trains or noise bursts, and it is not clear how to generalize these findings to speech sounds. Our system has implicitly learned the characterization of an "onset" that can provide precise localization.

5. CONCLUSIONS

This paper described a simple, practical method for improving audio source localization. We have demonstrated that the precision information provided by our technique reduces localization error compared to other, simpler techniques. In addition, the learned mappings are consistent with the precedence effect in that they are sensitive to sudden increases in audio energy. While it is impossible for the simple model we have learned to model all of the subtleties of the precedence effect, the similarities are encouraging. Future work will consist of relaxing the linear-Gaussian assumption implied by our use of FIR filters. While our linear FIR model is adequate to capture simple relationships between the spectrogram and localization precision, richer models should allow us to make use of more of the structure of human speech.

Thanks to John Fisher and Michael Siracusa for helpful discussions in the development of this work. This research was carried out in the Vision Interface Group, which is supported in part by DARPA and Project Oxygen.

6. REFERENCES

[1] Michael S. Brandstein and Darren Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Robust localization in reverberant rooms, Springer, 2001.

[2] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, 2000.

[3] Steven George Goodridge, *Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer Interaction*, Ph.D. thesis, North Carolina State University, 1997.

[4] Robert Eiichi Irie, "Robust sound localization : an application of an auditory perception system for a humanoid robot," M.S. thesis, Massachusetts Institute of Technology, 1995.

[5] Ehud Ben-Reuven and Yoram Singer, "Discriminative binaural sound localization," in *Advances in Neural Information Processing Systems 15*, S. Thrun S. Becker and K. Obermayer, Eds., pp. 1229–1236. MIT Press, Cambridge, MA, 2003.

[6] Ruth Y. Litovsky, H. Steven Colburn, William A. Yost, and Sandra J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999.

[7] William A. Yost and George Gourevitch, Eds., *Directional Hearing*, chapter The precedence effect, Springer-Verlag, 1987.

[8] George Christopher Stecker, *Observer weighting in sound localization*, Ph.D. thesis, University of California at Berkeley, 2000.

[9] Jont B. Allen and David A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[10] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 3rd edition, 1996.