LEARNING THE PRECEDENCE EFFECT: INITIAL REAL-WORLD TESTS

Kevin Wilson

Computer Science and Artificial Intelligence Lab, MIT
32 Vassar St.
Cambridge, MA 02139, USA
kwilson@csail.mit.edu

ABSTRACT

Audio source localization in reverberant environments has proved 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 give an overview of a system that we have designed that exploits these cues by learning a mapping from reverberated signal spectrograms to localization precision. We then describe initial tests of the system that demonstrate improved source localization on real audio data using the generalized cross-correlation (GCC) framework. We also relate the system's learned mappings to the well-known precedence effect from psychoacoustic studies.

1. INTRODUCTION

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

To allow for source motion, most practical localization systems compute localization cues based on short time segments of a few tens of milliseconds and combine these individual localization cues across time using a source motion model. In such a system, there are two broad areas where improvements can be made. The first is the low-level cues themselves, and the second is the means by which the cues are combined. Our system, first described in [3] and refined in [4], focuses on the latter area, learning an improved uncertainty model for the low-level cues that allows for improved fusion across frequency and time. We use cues from the reverberated audio to predict the uncertainty of localization cues derived from small time-frequency regions of the microphone array input.

This paper presents our initial tests of the system on real-world audio and elucidates a theoretical justification for our choice of generalized cross-correlation (GCC) weighting function.

Section 2 reviews related work in TDOA estimation and the psychoacoustics of the precedence effect. Section 3 describes our method for learning audio cues. Section 4 describes the results of our technique in a real reverberant environment and discusses the structure of our learned mappings as they relate to the precedence effect.

2. BACKGROUND

Our technique takes inspiration from the psychoacoustics literature on the precedence effect to generate a weighting function for a generalized cross-correlation-based source localizer. In this section, we review relevant work in these subjects.

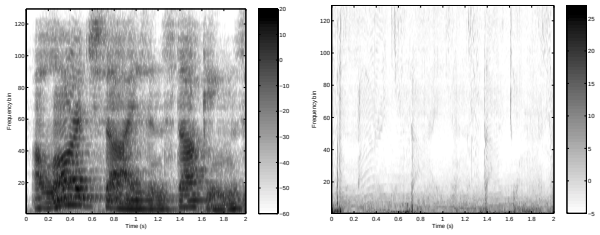
2.1. Array processing for source localization

In [1], DiBiase et al. review much of the work relevant to microphone arrays. They taxonomize source localization techniques into three groups – steered beamformer-based locators, high-resolution spectral estimation-based locators, and TDOA-based locators. Spectral estimation-based locators, while capable of high-resolution localization under ideal conditions, tend to be sensitive to modelling errors and also computationally expensive, which limits their use in practice. While in general steered-beamformer-based techniques and TDOA-based techniques differ, they are equivalent for the special case of a two element array, which is the case that we focus on in this paper. Therefore, we focus on TDOA-based techniques in the remainder of this section.

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. This technique performs well in anechoic environments, but performance degrades rapidly with increasing reverberation. Knapp and Carter [5] analyzed the generalized cross-correlation (GCC) framework, in which a frequency-dependent weighting is applied to reduce the effects of noise. [5] also derived an ML weighting for GCC that requires knowledge of the signal-to-noise ratio (SNR). Because the SNR is often unknown, the phase transform (PHAT) weighting, which simply whitens the microphone signals and works reasonably well in practice, is a popular alternative. In reverberant environments in particular, the PHAT weighting has been found to work well, and [6] showed that the PHAT weighting approximates the optimal weighting for stationary signals in noise-free reverberant environments. The intuitive justification for this technique is that no single frequency dominates, and that the effects of reverberation cancel out when averaged over many frequencies. Our technique defines a new GCC weighting that is a function of the reverberated speech spectrogram.

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



(a) Speech spectrogram (b) Localization precision map

Figure 1: Empirical justification for the precedence effect. Figure 1(a) is a spectrogram of the reverberant speech (a male voice saying "A large size in stockings...") received at one of the microphones in the array. Figure 1(b) is the corresponding map of the empirical localization precision (in dB) for each time-frequency bin. Sudden onsets in the spectrogram (a), such as those at 0.07, 0.7, and 1.4 seconds, correspond to time-frequency regions with high localization precision in (b). This figure was generated using simulated room reverberation as described in [4].

apparent location of a sound is influenced most strongly by the localization cues from the initial onset of the sound [7, 8]. For example, when human listeners report the location of a rapid sequence of clicks, they tend to report the location of the initial click even if later clicks in the sequence came from other directions [9]. 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.

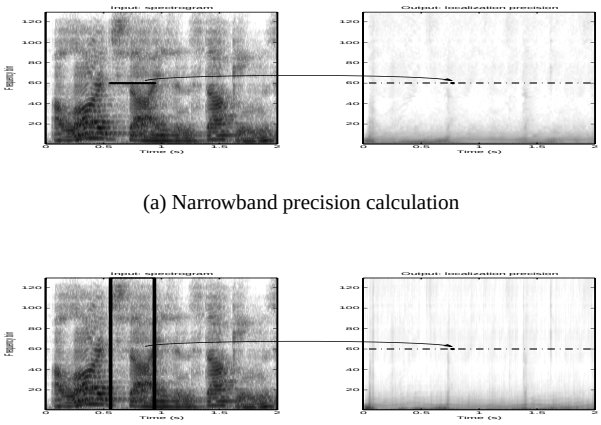
In [8], Zurek proposed a high-level conceptual model of the precedence effect without precisely specifying the details of the model. He modeled the precedence effect as a time-dependent weighting of raw localization cues. Specifically, his weighting took the raw audio as input and consisted of an "onset detector" with output generated by an inhibition function. In the next section, we describe a specific implementation of a model similar to Zurek's.

3. SYSTEM OVERVIEW

Our goal is to learn cues observable in the reverberated audio that indicate the reliability of associated localization cues. Specifically, we learn a mapping between the audio spectrogram and the localization precision, which we define to be the reciprocal of the empirical localization error variance. 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. Complete details of the system are in [4]; here we present an overview.

3.1. Corpus generation

To train the system, we collect a corpus of speech from known locations with a two-element microphone array. We then compute spectrograms of $y_n(i,t)$, with $i \in \{1,2\}$ representing the i^{th} microphone signal and with window size N_w, overlap N_o, and FFT length N_f, yielding complex spectrograms $s_n(i,u,f)$, where frame index u replaces the time index t, and frequency index f is added. We then calculate the cross-power spectrum phase (the frequency-domain equivalent of cross-correlation), $\theta_n(u,f)$, for each frame and frequency bin. Finally, we calculate $e(u,f) = (\theta_n(u,f) - \theta_{n_{true}}(u,f))^2$, the localization (phase) error variance, and $prec(u,f) = -10 * log_{10}(e(u,f))$, the localization precision (in dB).



(a) Narrowband precision calculation

(b) Broadband precision calculation

Figure 2: An illustration of the narrowband and broadband mappings for frequency bin 60. In 2(a) an FIR filter estimates the localization precision as a function of spectrogram bin 60. In 2(b) an FIR filter estimates the localization precision as a function of all spectrogram bins.

In [3] and [4], we used the image method to simulate room reverberation for testing and training. This allowed us to obtain good empirical localization error estimates by simulating many realizations of different source-microphone configurations within the same room. In this paper, we test and train on real data using the same procedure, but because of the difficulty of time-aligning multiple realizations, we treat our entire training corpus as a single realization. Thus, our training data is noisier than the simulated data from [3] or [4].

By calculating only these variances without cross-covariances we implicitly assume that errors in different time-frequency regions are uncorrelated. Although this is not strictly true, this assumption seems to work well in practice.

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 schematically 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 spec-

trogram. This case is shown schematically in Figure 2(b). We choose to examine the narrowband case because, for the case of stationary signals (and under the assumption of large spectrogram windows), each frequency band is uncorrelated with all other frequency bands, and thus the narrowband mapping should be sufficient in this case. Although 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 $\mathbf{z_f} = \mathbf{A_f b_f}$, $f \in \{1 \dots N_f\}$ where

$$\mathbf{z_f} = \left(\dots \; prec(u,f) \; prec(u+1,f) \; \dots \right)^{\mathsf{T}}$$

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

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

A similar system of equations can be solved to find the broadband filters. See [4] for details. For both types of mapping, we solve these systems using ridge regression by minimizing

$$||\mathbf{z_f} - \mathbf{A_f b_f}||^2 + \lambda ||\mathbf{b_f}||^2 \quad (2)$$

with respect to $\mathbf{b_f}$. The regularizing parameter λ is set through cross validation.

3.2. Applying the filters

We apply $\mathbf{b_f}$ to spectrogram $s_n(1,u,f)$ yielding $prec_{est}(u,f)$. We then use this estimated precision to create a GCC weighting for each frame. As defined in [5], a weighting, $\Psi(f)$ is applied to the cross-power spectrum of the two microphone signals before applying the inverse Fourier transform and locating the peak of this cross-correlation waveform. For example, the weighting for the phase transform is $\Psi(f) = 1/|G_{x_1x_2}(f)|$, where $G_{x_1x_2}$ is the cross-power spectrum of the two microphone signals. This weighting whitens the signals before cross-correlation. We define a weighting function based on our precision estimates as

$$\Psi(u,f) = \frac{prec_{est}(u,f)}{|G_{x_1x_2}(u,f)|} \quad (3)$$

As shown in [5], equation 46, this weighting is approximately equal to the ML GCC weighting. Although [5] assumes Gaussian signals and uncorrelated noise, it has become a standard benchmark when sufficient information is available to calculate it. It is a subject of future work to determine if these assumptions are necessary to prove the optimality of this localization precision-based ML weighting. Note that the phase transform is equivalent to setting $prec_{est}(u,f) = 1$.

When applying this technique to localization, the only computational costs (beyond the basic TDOA calculations) are of 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 strongly on the detailed structure of the reverberation, our technique is robust to changes in location in the room.

Room	Speaker	PHAT error	Narrow error	Broad error	Prop. error
Training	Train male	70	8	9	13
	Other male	66	8	10	9
	Female	71	7	8	7
Testing	Train male	90	12	14	21
	Other male	86	7	13	10
	Female	81	8	17	9

Table 1: Results within the training and testing rooms for the described weightings. Errors are root-mean-square (RMS) errors in μs. Before calculating RMS error, outlier estimates (error ≥ 200 μs) were removed.

4. RESULTS

In this evaluation, we use audio sampled at 8 kHz from two cardioid electret microphones spaced 37 cm apart, and we use a spectrogram with $N_w = 150$ and $N_o = 120$. We set our FFT size equal to 256. Thus, the frame rate for our spectrogram and for our TDOA estimates is 267 frames per second. We choose these parameters to be able to capture effects on the time scale at which the precedence effect has been observed, on the order of a few milliseconds. We use a total of 18 minutes of audio for training, collected over four different source-microphone configurations in a single room. The audio is a subset of the Harvard sentences [11] spoken by a single male and played through a desktop computer speaker. The room is rectangular and is 4m $\times$ 7m $\times$ 2.8m. The source-microphone distance ranged from 1.8m to 3m.

Our test data consists of audio from three speakers: the male speaker from the training set, a different male speaker, and a female speaker. None of the utterances used for testing were in the training set. Testing was done in two rooms: the room used for training and a larger, irregularly shaped room with approximately 1.5 times the volume of the training room. For test results in the training room, the source and microphone locations were different than those used for training.

For both testing and training, the primary sources of additive noise were computer fans and other ventilation sounds.

4.1. Localization results

Table 1 shows the decrease in localization error achieved by our technique on real data. The four columns of numbers are RMS error results (in μs) for four different GCC weighting functions. "PHAT" is the phase transform weighting function described in [5]. "Narrow" and "Broad" are the narrowband and broadband mappings described above. "Prop." is a simple special case of the narrowband filter using only one tap. This "proportional" mapping could express the simple relationship in which localization cues are weighted proportionally to the local signal power, but it cannot capture more complicated relationships. In all cases, our learned filter GCC weightings outperform the PHAT weighting.

Performance is good for the speaker from the training set and in the training room. This is to be expected since this corresponds closely to the training setup. Even in this simple case, though, we have generalized to new source and microphone locations, which

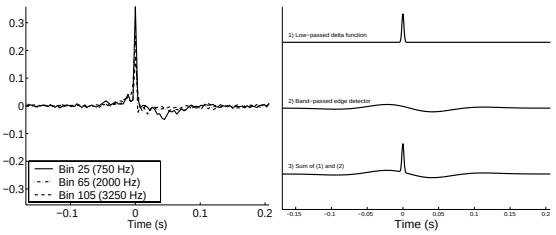


Figure 3: Narrowband filters. Left shows a representative subset of the learned filters. Right shows a schematic decomposition of the learned filters. Each of the narrowband filters on the left can be viewed as a linear combination of a low-pass filtered impulse (top) with a band-pass filtered edge detector (middle). The bottom curve shows the linear combination of the top two curves, which is qualitatively similar to the filter for bin 25.

we could not do if we were modelling the fine structure of the reverberation. The remaining results, with different speakers and/or different acoustic environments, show that our technique is robust to these changes.

Although "Narrow," "Broad," and "Prop." all perform well, "Narrow" performs best overall. This is somewhat surprising since in simulated experiments in [3] and [4], "Broad" performed best. A possible explanation for this is that "Broad" has many more parameters and thus can easily overfit the training data, while "Prop." has few parameters and may underfit. The training set of real data used in this paper was smaller than the training set of simulated data used previously, which would put the broadband mapping at a relative disadvantage. More experiments are necessary to fully explore this issue.

4.2. Relationship to the precedence effect

The left side of Figure 3 shows the FIR filters for a representative subset of the filter bands. In all three cases, but particularly for bin 25, the filter is approximately a superposition of a low-passed delta function and a band-passed edge-detector, as depicted schematically on the right of Figure 3. 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 correlated with 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 earlier work [3, 4], the edge-detector component was more prominent because the reverberation was stronger relative to the additive background noise.

Our results are consistent with the precedence effect, but they go beyond that by learning structure that is specific to the speech signal itself. For example, 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 on real audio data compared to the popular PHAT GCC technique. 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, which should allow us to make use of a wider range of audio cues in varied acoustical environments.

Thanks to Trevor Darrell, 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] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Microphone arrays: Signal processing techniques and applications," M. S. Brandstein and D. Ward, Eds. Springer, 2001, ch. Robust localization in reverberant rooms.

[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] K. Wilson and T. Darrell, "Improving audio source localization by learning the precedence effect," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[4] ——, "Learning the precedence effect in the generalized cross-correlation framework," *In submission*, 2005.

[5] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[6] T. Gustafsson, B. D. Rao, and M. Trivedi, "Source localization in reverberant environments: Modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 791–803, 2003.

[7] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 1633–1654, 1999. [Online]. Available: http://link.aip.org/link/?JAS/106/1633/1

[8] P. M. Zurek, "Directional hearing," W. A. Yost and G. Gourevitch, Eds. Springer-Verlag, 1987, ch. The precedence effect.

[9] G. C. Stecker, "Observer weighting in sound localization," Ph.D. dissertation, University of California at Berkeley, 2000.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Johns Hopkins University Press, 1996.

[11] J. P. Egan, "Articulation testing methods," *Laryngoscope*, vol. 58, pp. 955–991, 1948.