

Learning a Precedence Effect-Like Weighting Function for the Generalized Cross-Correlation Framework

Kevin W. Wilson, *Student Member, IEEE*, and Trevor Darrell, *Member, IEEE*

Abstract—Speech source localization in reverberant environments has proved difficult for automated microphone array systems. Because of its nonstationary nature, certain features observable in the reverberant speech signal, such as sudden increases in audio energy, provide cues to indicate time–frequency regions that are particularly useful for audio localization. We exploit these cues by learning a mapping from reverberated signal spectrograms to localization precision using ridge regression. Using the learned mappings in the generalized cross-correlation framework, we demonstrate improved localization performance. Additionally, the resulting mappings exhibit behavior consistent with the well-known precedence effect from psychoacoustic studies.

Index Terms—Acoustic arrays, array signal processing, delay estimation, direction of arrival estimation, speech processing.

I. 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. This paper 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. 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 as our low-level localization cues.

Our paper makes three contributions. First, we devise a method that uses recorded speech and simulated reverberation to generate a corpus of reverberated speech and the associated error for TDOA estimates made from this reverberated speech. Second, we use this corpus to learn mappings from the reverberated speech to a measure of TDOA 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 approximate the maximum likelihood (ML) weighting, we believe that we are the first to learn such a mapping from a training corpus. Our work in [5] introduced the idea of learning a mapping between spectrograms and localization precision but did not incorporate this mapping into the generalized cross-correlation framework.

Section II reviews related work in TDOA estimation and the psychoacoustics of the precedence effect. Section III describes our novel method of generating a training corpus and our method for learning audio cues. Section IV describes the results of our technique in a simulated reverberant environment and discusses the structure of our learned mappings as they relate to the precedence effect.

II. 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.

A. Array Processing for Source Localization

DiBiase *et al.* [1] 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 modeling 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. Therefore, we focus on TDOA-based techniques in the remainder of this section. (By focusing on TDOA-based techniques and a two element array,

Manuscript received January 31, 2005; revised September 29, 2005. This work was supported in part by the Defense Advanced Research Projects Agency and Project Oxygen and carried out in the Vision Interface Group. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Malcolm Slaney.

The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: kwilson@csail.mit.edu; trevor@csail.mit.edu).

Digital Object Identifier 10.1109/TASL.2006.872601

we also limit the localization information that we can obtain. In particular, we cannot estimate source range, and source angle is only known to within a “cone of confusion” that is symmetric about the array axis.)

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 at which the maximum of the cross-correlation signal occurs is assumed to be the TDOA. This technique performs well in anechoic environments, but performance degrades rapidly with increasing reverberation. Knapp and Carter [6] analyzed the generalized cross-correlation (GCC) framework, in which a frequency-dependent weighting is applied to reduce the effects of noise. [6] also derived an ML weighting for GCC that requires knowledge of the signal-to-noise ratio (SNR). (Here and throughout the paper, we use the term SNR to refer to the power ratio between the target source’s direct path signal and all other audio, including interfering sources and reflections of the target source.) Because the SNR is often unknown, the phase transform (PHAT) weighting, which simply whitens the microphone signals, is a popular alternative. In reverberant environments in particular, the PHAT weighting has been found to work well in practice, and [7] 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.

Attempts to improve localization performance fall into two broad categories—some systems attempt to build in robustness to reverberation at a very low level while others attempt to improve the way in which multiple localization cues are fused into a final location estimate.

In the first category are systems that attempt to use detailed models of the reverberation to undo its effects. For example, [8] 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. In general, [2] shows that the fine structure of the reverberation effects in a room can vary greatly and unpredictably over distances of tens of centimeters, so it is unclear how robust methods in this category 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 of techniques that improve cue fusion, and it provides a principled way of fusing cues based on mappings learned from a training corpus.

B. Precedence Effect

The precedence effect, also known as the “law of the first wavefront,” is the psychoacoustic effect in which the apparent

location of a sound is influenced most strongly by the localization cues from the initial onset of the sound [9], [10]. 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 correlated reflections, initial onsets will tend to be less corrupted by reverberation than subsequent sounds. The generality of this argument suggests that other animals should also exhibit the precedence effect, and evidence for the effect has been found in cats, dogs, rats, owls, and crickets [9].

Although the basic purpose of the precedence effect seems straightforward, the details are not clear. The notion of an “onset” is imprecise, although progress has been made in [11] in determining the time scales over which the precedence effect operates for click trains, and [12] shows the effect of onset duration on the ability to localize narrowband sounds. In addition, most studies have focused on stimuli such as click trains or noise bursts, and it is not obvious how to apply their findings to more natural sounds. For example, the effect is strongest in click pairs for inter-click intervals of roughly 2–10 ms [9]. Other effects dominate at shorter or longer time scales.

Studies on human infants and young puppies (reviewed in [9]) found no evidence of the precedence effect, and studies on young children have found the effect to be much smaller. Together with the stronger effects found in adults, this suggests that the precedence effect may be learned during childhood, although maturation of neural pathways, even in the absence of direct experience in reverberant environments, could also cause this gradual development of the effect. 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.

There is also evidence of adaptation over relatively short time scales. For example, in the “Clifton effect” [13], the precedence effect can be temporarily suppressed by suddenly swapping the locations of the leading and lagging clicks in a click-pair experiment. This is not the time scale at which we hope to apply learning, but it is an important consideration in psychoacoustically faithful models of the precedence effect.

A number of computational models of the precedence effect have been proposed. In [10], 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. Zurek’s high-level model was subsequently implemented and evaluated by Martin [14].

Lindemann [15], [16] presents a cross-correlation-based model of auditory lateralization, subsequently extended by Gaik [17], that includes an inhibition component that can model many aspects of the precedence effect. Lindemann’s model has many parameters whose values were chosen to accurately model human localization performance. Huang *et*

al. [18] present a more engineering-oriented model of the precedence effect and apply it to source localization. However, their approach makes an all-or-none decision about each localization cue and bases time delay estimates on differences between zero-crossing times instead of finding the maximum of a cross-correlation function. Recently, Faller and Merimaa [19] presented a model that uses estimated interaural coherence values to predict which time instants in a reverberated signal contain the best localization cues. They model many of the aspects of the precedence effect using these interaural coherence values, but their model does not explain why some steady-state sounds with high coherence are suppressed or why sounds originating in the median sagittal plane, which are perfectly coherent, can still elicit the precedence effect as shown in [20].

Our model can be viewed as a specific implementation of a model similar to Zurek's. However, our goal is not to faithfully model the human auditory system but to find a weighting function for the GCC framework that will accurately localize speech in reverberant environments. Because of this difference in approach, we do not incorporate elements such as psychoacoustically inspired filter banks or neural transduction models, and we do not try to model details such as the Clifton effect. Instead we focus on predicting the reliability of localization cues derived from a simple spectrogram representation. In comparison to other approaches, our approach relates directly to the GCC framework, which is the optimal TDOA estimator (under a set of assumptions enumerated in [6]) and provides a principled way to integrate localization cues across time and frequency. In contrast to Faller and Merimaa, who make use of interaural coherence, we predict localization precision based on solely monaural cues. It may help to incorporate interaural coherence cues into our system in the future, but essentially monaural cues must be able to trigger the precedence effect since the precedence effect is still observed for localization in the median sagittal plane [20].

III. METHODS

Our goal is to learn cues observable in the reverberated audio that indicate the reliability of associated localization cues. Specifically, we learn an association between the audio spectrogram and the localization precision, which we define to be the reciprocal of the empirical TDOA mean-squared 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 Fig. 1. We then compute a set of filters that estimate the localization precision from the spectrogram representation of the reverberated audio.

A. Corpus Generation

We generate the training corpus by using the image method of reverberation modeling [21] to simulate rooms 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

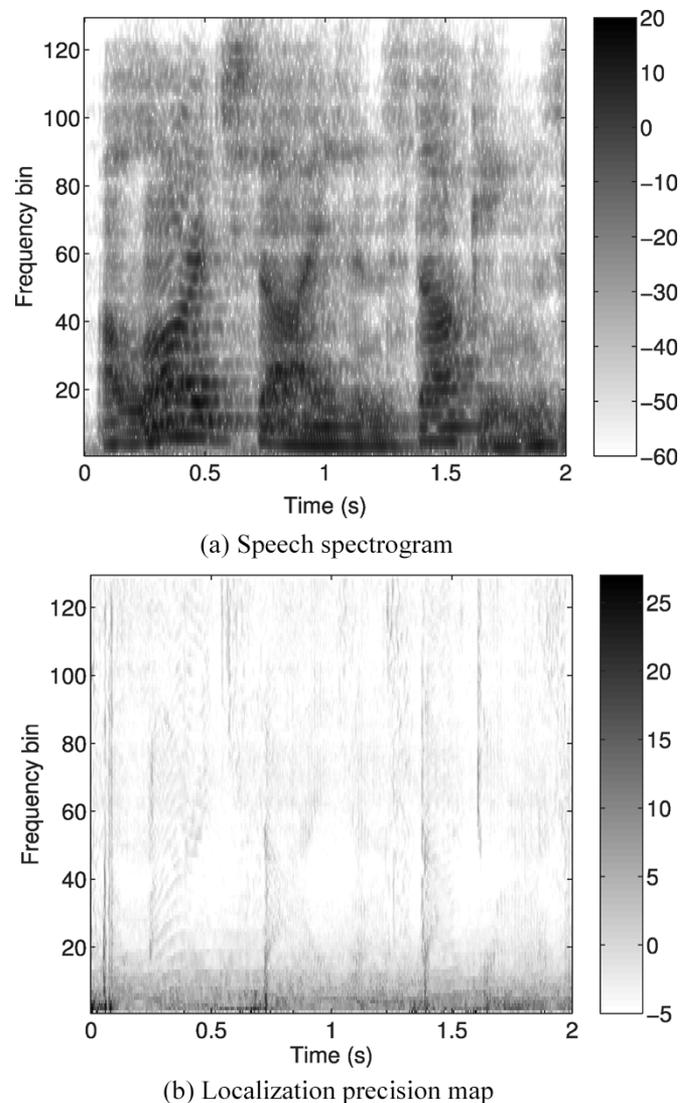


Fig. 1. Empirical justification for the precedence effect. (a) Spectrogram of the reverberant speech (a male voice saying “A large size in stockings . . .”) received at one of the microphones in the array. (b) Corresponding map of the empirical localization precision (in decibels) for each time–frequency bin. Sudden onsets in the spectrogram (a), such as those at 0.07, 0.7, and 1.4 s, correspond to time–frequency regions with high localization precision in (b).

such as computer fans and ventilation systems. We synthesize N_r realizations of the utterance, each with the speech source and microphones in random locations in a randomly generated room, and calculate the empirical localization precision over all realizations.

More formally, we start with a single speech signal, $x(t)$, and randomly generate N_r simulated room configurations. We represent these room 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 a noise signal $z_n(i, t)$ yields $y_n(i, t)$, a set of reverberated speech signals. We then compute spectrograms of $y_n(i, t)$ with window size N_w , overlap N_o , and fast Fourier transform (FFT) length N_f , yielding

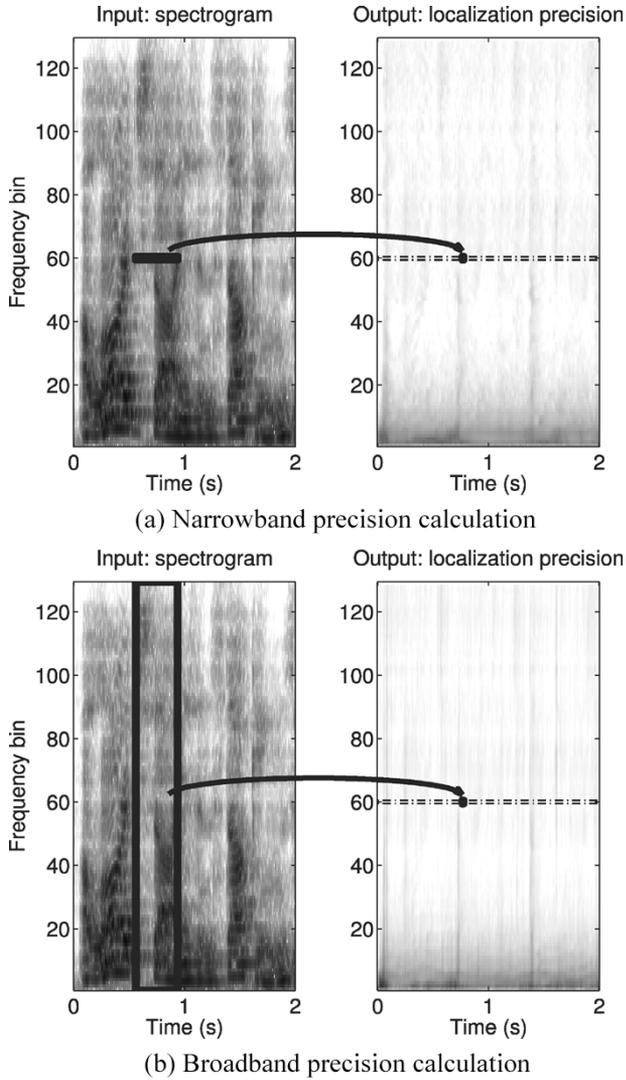


Fig. 2. Illustration of the narrowband and broadband mappings for frequency band 60. (a) An FIR filter estimates the localization precision as a function of spectrogram bin 60. (b) An FIR filter estimates the localization precision as a function of all spectrogram bins.

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 performing cross-correlation), $\theta_n(u, f)$, for each frame and frequency bin. Finally, we calculate $e(u, f) = (1/N_r) \sum_{n=1}^{N_r} (\theta_n(u, f) - \theta_{n_{true}}(u, f))^2$, the localization (wrapped phase) error variance, and $prec(u, f) = -10 * \log_{10}(e(u, f))$, the localization precision (in decibels). Fig. 3(a) shows a block diagram describing these calculations.

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

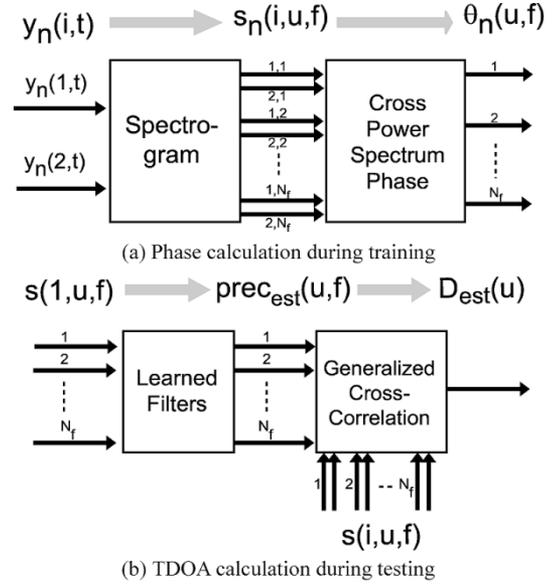


Fig. 3. (a) Procedure for calculating the cross-power spectrum phase used during training. (b) Procedure for using our estimated precision map to calculate TDOA during testing.

B. Filter Learning

We then use ridge regression [22] to learn finite-impulse response (FIR) filters that estimate the localization precision (in decibels) from the reverberated spectrogram (in decibels). 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 Fig. 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 shown schematically in Fig. 2(b). We choose to examine the narrowband case because, for the case of stationary signals (and under the assumption of spectrogram windows that are much larger than the coherence time of the signal), each frequency band is uncorrelated with all other frequency bands, and thus the narrowband mapping should be sufficient in this case. Although speech is non-stationary, this narrowband mapping provides a useful baseline against which to compare. Additionally, in [12], the precedence effect was demonstrated with narrowband sounds, where the onset rate of a sinusoidal tone affected the ability to localize that tone, which is exactly the relationship that our narrowband mapping can express. 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 anti-causal length l_{ac} , we solve N_f regularized linear least-squares problems of the form $\mathbf{z}_f = \mathbf{A}_f \mathbf{b}_f$, $f \in \{1 \dots N_f\}$ as shown in (1) at the bottom of the page, 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 form $\mathbf{z}_f = \mathbf{A}_f \mathbf{b}_f$, as shown in (2) at the bottom of the page, and \mathbf{b}_f is an FIR filter with $(l_c + l_{ac} + 1) * N_f$ taps stacked with a dc component. For both types of mapping, we solve these systems using ridge regression by minimizing

$$\|\mathbf{z}_f - \mathbf{A}_f \mathbf{b}_f\|^2 + \lambda \|\mathbf{b}_f\|^2 \quad (3)$$

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

C. Applying the Filters

We apply filters \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 [6], 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 GCC weighting for the phase transform is $\Psi(f) = 1/|G_{x_1 x_2}(f)|$, where $G_{x_1 x_2}$ 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

$$\Psi(u, f) = \frac{prec_{est}(u, f)}{|G_{x_1 x_2}(u, f)|}. \quad (4)$$

Thus, 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 the acoustic environment.

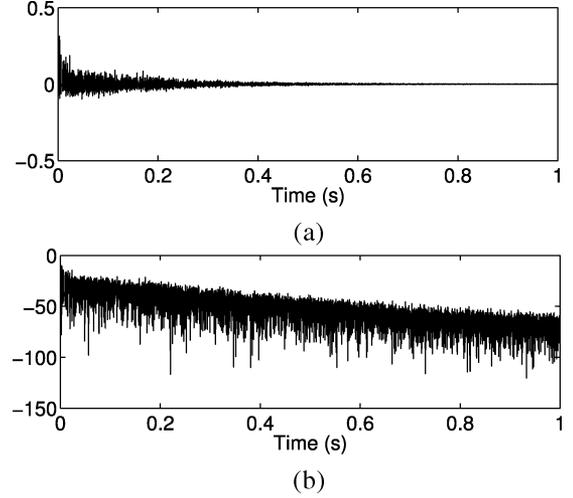


Fig. 4. Typical reverberant impulse response generated using the image method. This impulse response has a reverberation time of 1.5 s. (a) Typical reverberant impulse response. (b) Typical reverberant impulse response, log magnitude.

IV. RESULTS

In this evaluation, we use audio sampled at 8 kHz, 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 17 min of speech for training, and a separate 90 s of speech for testing. We simulate an ensemble of rooms ranging in size from $3 \times 6 \times 3$ m to $6 \times 9 \times 4$ m with wall acoustic reflection characteristics randomly chosen from a database of wall materials such as glass, plaster, plywood, and acoustic tiles. The reverberation times of the simulated rooms ranged from 0.3 to 2.5 s. Fig. 4 shows a typical reverberant impulse response with a reverberation time

$$\mathbf{z}_f = (\dots prec(u, f) prec(u+1, f) \dots)^\top$$

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

$$\mathbf{z}_f = (\dots prec(u, f) prec(u+1, f) \dots)^\top$$

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

TABLE I

TEST RESULTS FOR NEW UTTERANCES BY THE SPEAKER USED IN TRAINING FOR THE DESCRIBED WEIGHTINGS. $P(\text{error} \leq 50 \mu\text{s})$ IS THE PROBABILITY OF BEING WITHIN $50 \mu\text{s}$ OR 5° OF THE TRUE TDOA

Method	RMS TDOA error (μs)	RMS angular error (degrees)	$P(\text{error} \leq 50 \mu\text{s})$
True precision	64	6.3	0.90
Broadband	74	7.3	0.78
Narrowband	79	7.7	0.78
Proportional	92	9.0	0.75
PHAT	250	25.3	0.55

of 1.5 s. Source to microphone distances range from 1 to 4.5 m in our data. Our training corpus is a subset of the Harvard sentences [23].

Our training data consists of $N_r = 60$ sampled room configurations, and our performance results are averaged over $N_{\text{test}} = 12$ room configurations that do not appear in the training set.

A. Localization Results

Table I shows the decrease in localization error achieved by using our technique to generate a weighting function and localizing using GCC. For our first test, data generated from different utterances and in different rooms than any of the training data was synthesized. From this, we generated pairs of test spectrograms $s_{\text{test}}(i, u, f)$. The mappings learned according to the method in Section III-B were applied to $s_{\text{test}}(1, u, f)$, yielding an estimated localization precision map $\text{prec}_{\text{est}}(u, f)$. For each frame, $\text{prec}_{\text{est}}(u, f)$ is used to create a GCC weighting function as described in Section III-C.

Once we have calculated a delay estimate $d_{\text{test}}(u)$ for each frame, we can smooth the delay estimates across time using an appropriate source motion model. For the results presented here, we test on a stationary source and assume a simple model in which the source is stationary over 0.2-s intervals. To calculate a location over this interval, we compute generalized cross-correlation signals for each time frame in this interval, sum the cross-correlation signals, and then find the peak in this summed cross-correlation signal. Table I shows root-mean-square (rms) localization error achieved by each method for these TDOA estimates on our test set. The TDOA error in seconds is fundamentally what we measure, but we also compute a corresponding angular error since this will often be the most practically relevant performance metric. This angular error depends on the array geometry and source angle; numbers in the table assume a microphone spacing of 20 cm and a source directly broadside of the array.

Each row shows the performance of a different method of estimating precision information. The first row, “True precision,” shows localization 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 on a per-frame basis assuming that the error variance estimates from our training set are correct and our signal and noise are Gaussian and stationary with a coherence time that is short compared to the FFT length [6]. Although these assumptions are unlikely to be satisfied in practice, this is a useful baseline for comparison, and

TABLE II

RESULTS FOR A NEW SPEAKER. $P(\text{error} \leq 50 \mu\text{s})$ IS THE PROBABILITY OF BEING WITHIN $50 \mu\text{s}$ OR 5° OF THE TRUE TDOA

Method	RMS TDOA error (μs)	RMS angular error (degrees)	$P(\text{error} \leq 50 \mu\text{s})$
Broadband	87	8.5	0.74
Narrowband	88	8.7	0.74
Proportional	101	10.0	0.71
PHAT	225	22.6	0.56

TABLE III

RESULTS FOR UNUSUAL TEST SCENARIOS. ALL ERRORS ARE RMS TIME DELAY ERRORS IN MICROSECONDS. IN THE “NEAR WALL” SCENARIO, THE MICROPHONE ARRAY WAS WITHIN 10 cm OF THE WALL. IN THE “DISTANT SOURCE” SCENARIO, THE SOURCE-MICROPHONE DISTANCE WAS 4.5 m. IN THE “CLOSE SOURCE” SCENARIO, THE SOURCE-MICROPHONE DISTANCE WAS 1 m

Method	Near wall error (μs)	Distant source error (μs)	Close source error (μs)
True precision	33	141	12
Broadband	56	157	20
Narrowband	57	178	21
Proportional	60	193	27
PHAT	305	515	59

to the extent that our weightings underperform the true precision it is presumably due to their inability to perfectly reconstruct the true precision. “Broadband” and “Narrowband” are the mappings described in Section III-B. “Proportional” 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. “PHAT” is the phase transform and corresponds to uniformly weighting the localization cues in each time–frequency region (setting $\text{prec}_{\text{est}}(u, f) = 1$). In all cases, variants of our technique outperform the phase transform. The fact that our narrowband and broadband mappings outperform the proportional mapping shows that there is a practical benefit to using these richer mappings which are sensitive to energy distribution across time and frequency.

In addition to reducing localization error variance, our precision estimates also reduce the number of outliers. The rightmost column of Table I shows the probability, using each precision estimate, of being within $50 \mu\text{s}$ (5°) of the correct answer. These results are consistent with our error variance results and again show the effectiveness of our mappings.

The above results demonstrate the usefulness of our approach for speech signals drawn from the same distribution as our training data, but two important practical issues are the extent to which our learned mappings generalize to other speakers and the robustness of our approach to scenarios that were not well-represented in the training set. In Table II, we present results for an out-of-training-set speaker. In Table III, we present results from scenarios at the extremes of our training and testing parameter ranges. The results in both tables are similar to those from our initial tests, thus demonstrating that our mappings are general enough to handle some variation in source and environmental acoustics. Note that in Table III, the PHAT weighting performs more poorly than average for the “near wall” scenario, presumably because of a particularly

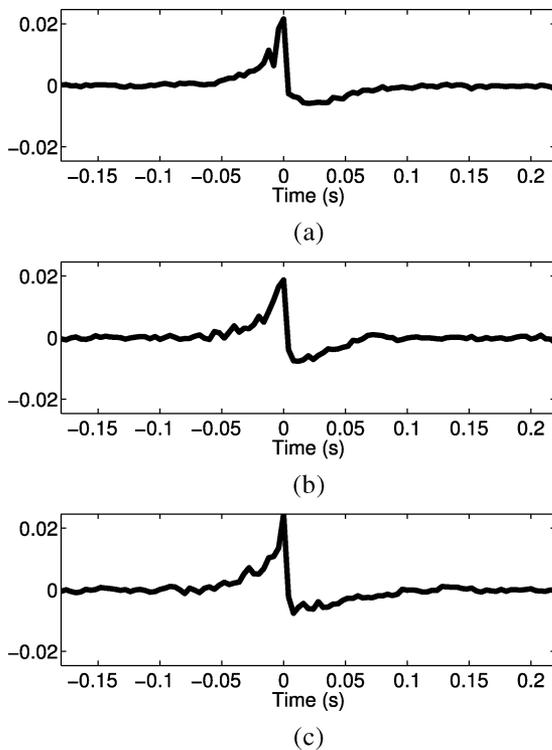


Fig. 5. Representative subset of the learned narrowband filters. (a) Freq. bin 30 (906 Hz). (b) Freq. bin 70 (2156 Hz). (c) Freq. bin 110 (3406 Hz).

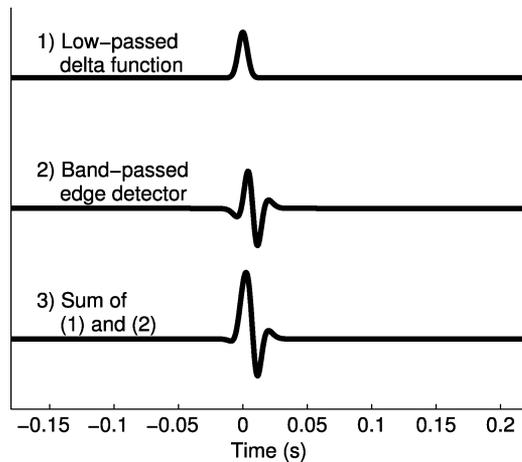


Fig. 6. Schematic decomposition of the learned filters. Each of the learned narrowband filters can be viewed as a linear combination of a lowpass filtered impulse (top) with a bandpass filtered edge detector (middle). The bottom curve shows the linear combination of the top two curves, which is qualitatively similar to the learned narrowband filters.

strong reflection from that wall, but our learned weightings yield good performance. Our learned weightings are also robust to changes in the direct-to-reverberant energy ratio, which varies greatly between the “distant source” and “close source” scenarios. In addition, our preliminary positive results on data recorded in real rooms, described in [24], suggest that our technique’s success is not limited to simulated environments.

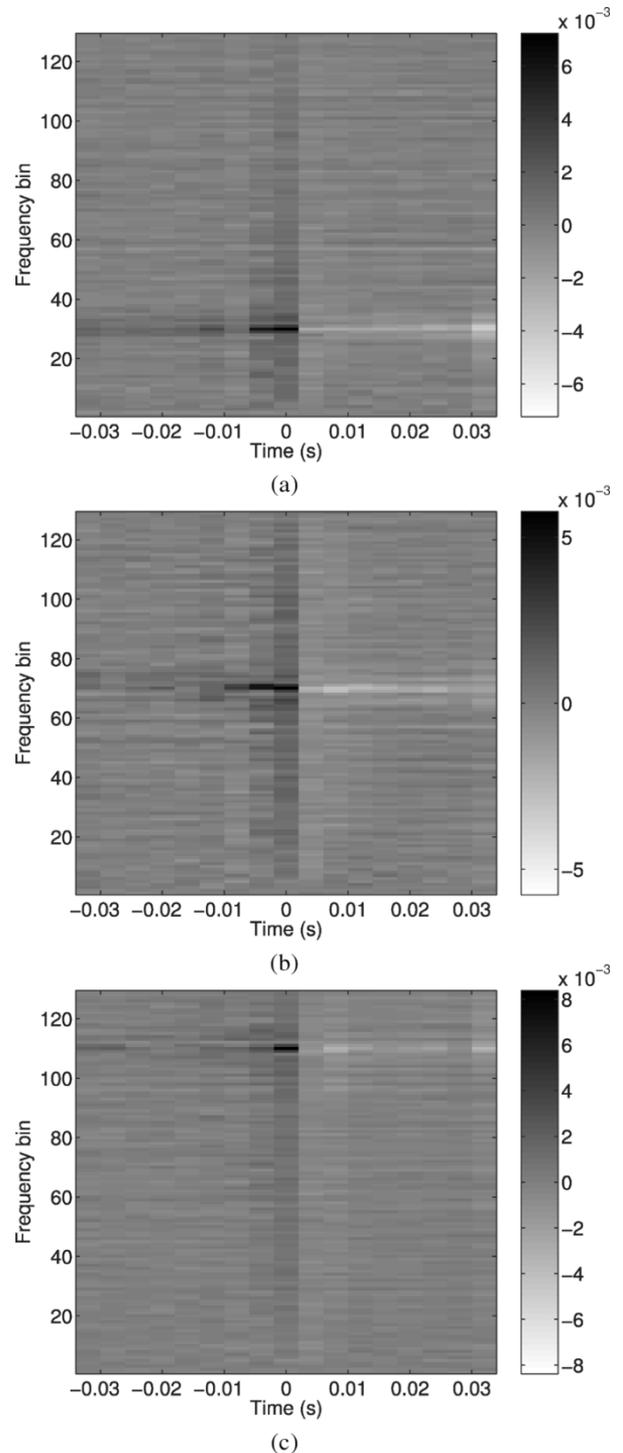


Fig. 7. Learned broadband filters for three representative filter bands. These filters have most of their energy in the frequency bin whose precision they are estimating, but there is some energy across all frequency bins, indicating that useful information is being integrated across frequency when calculating the optimal mapping. (a) Freq. bin 30 (906 Hz). (b) Freq. bin 70 (2156 Hz). (c) Freq. bin 110 (3406 Hz).

B. Relationship to the Precedence Effect

Fig. 5(a)–(c) shows the 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 Fig. 6. The low-passed delta function component indicates that louder sounds provide better localization cues since for a delta function impulse response a larger input (louder sound) will produce a proportionally larger output (higher-precision localization cue). This 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.

Our results are consistent with the precedence effect, and they also have learned 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 indicating that information is being integrated across all frequencies to calculate the optimal weighting, most obviously in Fig. 7(b). This is clearly useful since the broadband mapping performed better than the narrowband mapping. 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 over the range of acoustic environments on which we have trained.

V. CONCLUSION

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 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.

ACKNOWLEDGMENT

The authors would like to thank J. Fisher and M. Siracusa for helpful discussions in the development of this work. They would also like to thank the anonymous reviewers for their many comments, including their pointers to related work and suggestions for improvements to the experiments.

REFERENCES

- [1] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust localization in reverberant rooms,” in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds. New York: Springer, 2001.
- [2] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, “Equalization in an acoustic reverberant environment: robustness results,” *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 311–319, May 2000.
- [3] S. G. Goodridge, “Multimedia Sensor Fusion for Intelligent Camera Control and Human-Computer Interaction,” Ph.D. dissertation, North Carolina State Univ., Raleigh, 1997.
- [4] R. E. Irie, “Robust Sound Localization: An Application of an Auditory Perception System for a Humanoid Robot,” M.S. thesis, Mass. Inst. Technol., Cambridge, 1995.
- [5] K. Wilson and T. Darrell, “Improving audio source localization by learning the precedence effect,” in *IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Mar. 2005, vol. 4, pp. iv/1125–iv/1128.
- [6] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
- [7] T. Gustafsson, B. D. Rao, and M. Trivedi, “Source localization in reverberant environments: modeling and statistical analysis,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 791–803, Nov. 2003.
- [8] E. Ben-Reuven and Y. Singer, “Discriminative binaural sound localization,” in *Advances in Neural Information Processing Systems 15*, S. T. S. Becker and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 1229–1236.
- [9] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, “The precedence effect,” *The J. Acoust. Soc. Amer.* vol. 106, no. 4, pp. 1633–1654, 1999 [Online]. Available: <http://link.aip.org/link/?JAS/106/1633/1>
- [10] P. M. Zurek, “The precedence effect,” in *Directional Hearing*, W. A. Yost and G. Gourevitch, Eds. New York: Springer-Verlag, 1987.
- [11] G. C. Stecker, “Observer Weighting in Sound Localization,” Ph.D. dissertation, Univ. California, Berkeley, , 2000.
- [12] B. Rakerd and W. M. Hartmann, “Localization of sound in rooms, iii: Onset and duration effects,” *J. Acoust. Soc. Amer.* vol. 80, no. 6, pp. 1695–1706, 1986 [Online]. Available: <http://link.aip.org/link/?JAS/80/1695/1>
- [13] R. K. Clifton, “Breakdown of echo suppression in the precedence effect,” *J. Acoust. Soc. Amer.* vol. 82, no. 5, pp. 1834–1835, 1987 [Online]. Available: <http://link.aip.org/link/?JAS/82/1834/1>
- [14] K. D. Martin, “A Computational Model of Spatial Hearing,” M.S. thesis, Mass. Inst. Technol., Cambridge, 1995.
- [15] W. Lindemann, “Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals,” *J. Acoust. Soc. Amer.* vol. 80, no. 6, pp. 1608–1622, 1986 [Online]. Available: <http://link.aip.org/link/?JAS/80/1608/1>
- [16] ———, “Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front,” *J. Acoust. Soc. Amer.*, vol. 80, no. 6, pp. 1623–1630, 1986 [Online]. Available: <http://link.aip.org/link/?JAS/80/1623/1>
- [17] W. Gaik, “Combined evaluation of interaural time and intensity differences: psych oacoustic results and computer modeling,” *J. Acoust. Soc. Amer.*, vol. 94, no. 1, pp. 98–110, 1993 [Online]. Available: <http://link.aip.org/link/?JAS/94/98/1>
- [18] J. Huang, N. Ohnishi, and N. Sugie, “Sound localization in reverberant environment based on the model of the precedence effect,” *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, Aug. 1997.
- [19] C. Faller and J. Merimaa, “Source localization in complex listening situations: selection of binaural cues based on interaural coherence,” *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, 2004 [Online]. Available: <http://link.aip.org/link/?JAS/116/3075/1>
- [20] R. Y. Litovsky, B. Rakerd, T. C. Yin, and W. M. Hartmann, “Psychophysical and physiological evidence for a precedence effect in the median sagittal plane,” *J. Neurophysiol.*, vol. 77, no. 4, pp. 2223–2226, 1997 [Online]. Available: <http://jn.physiology.org/cgi/content/abstract/77/4/2223>
- [21] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979 [Online]. Available: <http://link.aip.org/link/?JAS/65/943/1>
- [22] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
- [23] J. P. Egan, “Articulation testing methods,” *Laryngoscope*, vol. 58, pp. 955–991, 1948.
- [24] K. Wilson, “Learning the precedence effect: Initial real-world tests,” in *IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, 2005, pp. 142–145.



Kevin Wilson (S'00) received the B.S. and M.Eng. degree in computer science and electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1999 and 2000, respectively. He is currently pursuing the Ph.D. degree in the Computer Science and Artificial Intelligence Laboratory at MIT.



Trevor Darrell (M'96) received the B.S.E. degree while working at the GRASP Robotics Laboratory at the University of Pennsylvania, Philadelphia, in 1988 and the S.M. and Ph.D. degrees from the Massachusetts Institute of Technology (MIT) Media Laboratory, Cambridge, in 1991 and 1996, respectively.

He is an Associate Professor of Electrical Engineering and Computer Science at MIT. He leads the Vision Interface Group at the Computer Science and Artificial Intelligence Laboratory. His interests include computer vision, interactive graphics, and machine learning. Prior to joining the faculty of MIT, he worked as a Member of the Research Staff at Interval Research, Palo Alto, CA, researching vision-based interface algorithms for consumer applications.