AUDIO INTERPOLATION

RICHARD RADKE[1] **AND SCOTT RICKARD**[2]

[1]*Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA*
rjradke@ecse.rpi.edu
[2]*Program in Applied and Computational Mathematics , Princeton University, Princeton, NJ 08540, USA*
srickard@princeton.edu

Using only the audio signals from two real microphones and the distance separating them, we synthesize the audio that would have been heard at any point along the line connecting the two microphones. The method is valid in anechoic environments. The interpolated audio can be calculated directly, with no need to estimate the number of sources present in the environment or to separate the sources from the received audio mixtures. However, additionally estimating the mixing parameters is shown to dramatically improve results for speech mixtures. Experimental results are presented, and sample sound files can be found on the authors' web site, http://www.ecse.rpi.edu/homepages/rjradke/pages/ainterp/ainterp.html.

INTRODUCTION

In this paper, our goal is to understand when there is enough information contained in the audio signals received at two microphones to produce the audio that would have actually been heard had a third microphone been present in the environment. We show that for anechoic environments, when the virtual microphone is located along the line connecting the two real microphones, the audio can be synthesized with no knowledge besides the distance between the two microphones.

We call our algorithm "audio interpolation" as an analogy to the term "view interpolation" [1, 2] from computer vision. View interpolation techniques use two or more real views of a scene to synthesize other, physically consistent views of the scene from the perspective of a virtual camera. By combining audio interpolation with view interpolation, we can obtain "virtual video" [3] that contains both images and sound.

Jourjine et al. [4] presented a novel method (the "DUET" algorithm) for blindly separating any number of sources using only two mixtures. The main assumption of the algorithm is that the sources are W-disjoint orthogonal, i.e. the supports of the windowed Fourier transforms of each pair $(s_i(t), s_j(t))$ of source signals are disjoint. This assumption has been shown to be true in an approximate sense for mixtures of multiple voices speaking simultaneously [5]. The mixing parameters of the sources are estimated by clustering ratios of the time-frequency representations of the mixtures. The estimates of the mixing parameters can then be used to partition the time-frequency representation of one mixture to recover the original sources. The technique is valid even when the number of sources is larger than the number of mixtures.

Our audio interpolation algorithm decouples into two parts. First, a method based on the DUET algorithm is used to blindly associate a physical location with each time-frequency point of the mixtures based on a model of anechoic mixing. Second, the time-frequency representations of the mixtures are altered to synthesize the virtual audio signals as they would have been heard at a third microphone placed along the line connecting the two microphones. We expect this research to have immediate applications in video conferencing and virtual video.

To our knowledge, audio interpolation is a previously unaddressed problem. We note that Slaney et al. proposed an algorithm called "audio morphing" [6], a method for automatically transitioning from one sound into another. While the intermediate signals may sound plausible, they do not correspond to sound produced by real underlying sources and microphones. This is in direct analogy to the computer graphics term "morphing", which produces intermediate images that correspond to no real physical objects.

1. MIXING MODEL AND SOURCE ASSUMPTIONS

We consider the measurements from a pair of monophonic, omnidirectional microphones, M_0 and M_1, in the presence of an unknown number of omnidirectional sources. We assume that only the direct path between each source and microphone is present, that is, the mixture is anechoic. The two mixtures can be expressed as

$$m_0(t) = \sum_{j=1}^{N} \frac{1}{d_{0j}} s_j\left(t - \frac{d_{0j}}{c}\right) \quad (1)$$

$$m_1(t) = \sum_{j=1}^{N} \frac{1}{d_{1j}} s_j\left(t - \frac{d_{1j}}{c}\right) \quad (2)$$

where $s_j(t)$, $j = 1, \ldots, N$, are the N sources, d_{kj} is the distance between the source j and microphone k, and c is the speed of propagation, in this case, the speed of sound ($\approx$ 343 m/s). We have assumed the source amplitudes decay as the inverse of distance traveled. We use

Δ to denote the distance separating M_0 and M_1; hence, $|d_{0j} - d_{1j}| \leq \Delta, \forall j$. We assume that the microphones are calibrated in the sense that the distance Δ between them is known. Figure 1 depicts the setup for source j. In Section 3 we will discuss how to obtain the signal for the "virtual" microphone M_α.

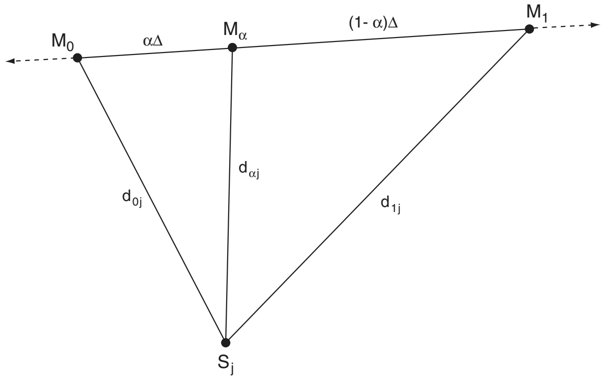

Figure 1: Microphone configuration.

We call two functions $s_1(t)$ and $s_2(t)$ **W-disjoint orthogonal** if, for a given a windowing function $W(t)$, the supports of the windowed Fourier transforms of $s_1(t)$ and $s_2(t)$ are disjoint [4]. The windowed Fourier transform of $s_j(t)$ is defined as

$$F^W(s_j(\cdot))(\omega,\tau) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t-\tau)s_j(t)e^{-i\omega t}dt, \tag{3}$$

which we will refer to as $\hat{s}_j(\omega,\tau)$ where appropriate. The W-disjoint orthogonality assumption can be stated concisely as

$$\hat{s}_1(\omega,\tau)\hat{s}_2(\omega,\tau) = 0, \forall \omega, \tau. \tag{4}$$

For speech signals, this condition has been shown to hold in an approximate sense [5].

When $W(t) = 1$, the following is a property of the Fourier transform:

$$F^W(s_j(\cdot - t_0))(\omega,\tau) = e^{-i\omega t_0} F^W(s_j(\cdot))(\omega,\tau). \tag{5}$$

We employ the narrowband assumption from array processing that (5) holds for all t_0 with $|t_0| \leq \frac{\Delta}{c}$, even when $W(t)$ has finite support [7].

In practice, rather than working with the continuous windowed Fourier transform, we use its discrete counterpart, $\hat{s}_j(k\omega_0, l\tau_0)$, $\forall k, l \in \mathbb{Z}$, where ω_0 and τ_0 are the frequency and time grid spacing parameters. It is well known that for any appropriately chosen window function, if ω_0 and τ_0 are small enough, $s_j(t)$ can be reconstructed from $\hat{s}_j(k\omega_0, l\tau_0)$, $\forall k, l \in \mathbb{Z}$. For more details, consult [8].

2. ESTIMATING PATH LENGTHS

As noted in [4], using the narrowband assumption, we can rewrite the model from (1) and (2) in the time-frequency domain as

$$\hat{m}_0(\omega,\tau) = \sum_{j=1}^{N} \frac{1}{d_{0j}} e^{-i\omega \frac{d_{0j}}{c}} \hat{s}_j(\omega,\tau) \tag{6}$$

$$\hat{m}_1(\omega,\tau) = \sum_{j=1}^{N} \frac{1}{d_{1j}} e^{-i\omega \frac{d_{1j}}{c}} \hat{s}_j(\omega,\tau) \tag{7}$$

Assuming the sources are pairwise W-disjoint orthogonal, at most one of the N sources will be non-zero for a given (ω,τ). We denote the index of this source as J, suppressing the dependence on (ω,τ). Thus

$$\hat{m}_0(\omega,\tau) = \frac{1}{d_{0J}} e^{-i\omega \frac{d_{0J}}{c}} \hat{s}_J(\omega,\tau) \tag{8}$$

$$\hat{m}_1(\omega,\tau) = \frac{1}{d_{1J}} e^{-i\omega \frac{d_{1J}}{c}} \hat{s}_J(\omega,\tau) \tag{9}$$

The key observation is that the ratio of the time-frequency representations of the mixtures is a function of the mixing parameters only and does not depend on the sources, that is,

$$\frac{\hat{m}_0(\omega,\tau)}{\hat{m}_1(\omega,\tau)} = \frac{d_{1J}}{d_{0J}} e^{-i\omega(d_{0J}-d_{1J})/c} \tag{10}$$

Therefore, we can calculate the ratio of the distances, $\frac{d_{1J}}{d_{0J}}$, and difference of the distances, $d_{1J} - d_{0J}$, associated with the source active at (ω,τ) as

$$\rho(\omega,\tau) = \left| \frac{\hat{m}_0(\omega,\tau)}{\hat{m}_1(\omega,\tau)} \right| \tag{11}$$

$$\delta(\omega,\tau) = \frac{c}{\omega} \measuredangle \frac{\hat{m}_0(\omega,\tau)}{\hat{m}_1(\omega,\tau)} \tag{12}$$

where $\measuredangle ae^{i\phi} = \phi$, $-\pi < \phi \leq \pi$. Note that the accurate calculation of $\delta(\omega,\tau)$ requires that

$$|d_{1J} - d_{0J}| < c\pi/\omega \tag{13}$$

to avoid phase wrap-around problems. This can be guaranteed provided $\Delta < \frac{171.5}{f_{\max}}$ meters where $f_{\max}$ is the maximum frequency component of the sources. For example, to guarantee correct estimation of $|d_{1J} - d_{0J}|$ for all frequencies up to 1000 Hz, we would require $\Delta \leq$ 17.15 cm.

From the ratio and the difference of the source to microphone distances, the actual source to microphone distances for the source active at (ω,τ) can be calculated as

$$d_{0J} = \frac{\delta(\omega,\tau)}{\rho(\omega,\tau) - 1} \tag{14}$$

$$d_{1J} = \frac{\rho(\omega,\tau)\delta(\omega,\tau)}{\rho(\omega,\tau) - 1} \tag{15}$$

3. AUDIO INTERPOLATION

We are interested in synthesizing the signal that would have been received at a microphone M_α placed a fraction α of the way along the line connecting M_0 to M_1 (the "baseline"). From Figure 1 and the law of cosines, we can compute

$$d_{\alpha J} = \sqrt{\alpha(1-\alpha)\Delta^2 + (1-\alpha)d_{0J}^2 + \alpha d_{1J}^2} \qquad (16)$$

This formula is correct for any value of α, not just $\alpha \in [0,1]$. That is, the virtual microphone can range anywhere along the line through M_0 and M_1. This means that we may take the original microphones to be as close together as we like, which may be desirable for dealing with phase wrap-around considerations of (13). We can also see from (16) that knowledge of the microphone separation is necessary; the dependence on Δ is removed only in the trivial cases when the virtual microphone is at $\alpha = 0$ or $\alpha = 1$.

We note that the line through M_0 and M_1 is unique in that it is the only location where the sound from a virtual microphone can be synthesized from only two microphones and multiple unknown sources. Using the signals $m_0(t)$ and $m_1(t)$, each source can be located only up to a point on a circle orthogonal to the baseline. Three calibrated, non-collinear microphones can locate each source up to a pair of points; four calibrated, non-coplanar microphones can locate each source unambiguously. Thus, two calibrated microphones is really the only "interesting" case in which accurate virtual audio can be synthesized from incomplete information.

In the practical case, when W-disjoint orthogonality holds only approximately, we compute estimates of d_{0J} and d_{1J} given by the values in each time-frequency bin:

$$\begin{aligned} d_0(\omega,\tau) &= \frac{\delta(\omega,\tau)}{\rho(\omega,\tau)-1} && (17) \\ d_1(\omega,\tau) &= \frac{\rho(\omega,\tau)\delta(\omega,\tau)}{\rho(\omega,\tau)-1}. && (18) \end{aligned}$$

The corresponding estimate of $d_{\alpha J}$ is given by

$$d_\alpha(\omega,\tau) = \sqrt{\alpha(1-\alpha)\Delta^2 + (1-\alpha)d_0(\omega,\tau)^2 + \alpha d_1(\omega,\tau)^2} \qquad (19)$$

Given $d_\alpha(\omega,\tau)$, we can modify the time-frequency representation of either mixture to obtain an estimate of the time-frequency representation of the virtual signal measured at M_α:

$$\begin{aligned} \hat{m}_\alpha(\omega,\tau) &= \frac{d_0(\omega,\tau)}{d_\alpha(\omega,\tau)} e^{i\omega\left(\frac{d_0(\omega,\tau)}{c} - \frac{d_\alpha(\omega,\tau)}{c}\right)} \hat{m}_0(\omega,\tau) && (20) \\ &= \frac{d_1(\omega,\tau)}{d_\alpha(\omega,\tau)} e^{i\omega\left(\frac{d_1(\omega,\tau)}{c} - \frac{d_\alpha(\omega,\tau)}{c}\right)} \hat{m}_1(\omega,\tau) && (21) \end{aligned}$$

Converting $\hat{m}_\alpha(\omega,\tau)$ back into the time domain yields the interpolated virtual microphone signal.

We emphasize that while the development of the algorithm is based on the assumption of W-disjoint orthogonality, the actual implementation need not involve separating the sources from the mixtures. That is, the method can be implemented entirely by (11)-(12) and (17)-(21), without any estimate of the number of sources or the time-frequency bins where various sources are active. However, deviations from the W-disjoint orthogonality assumption cause errors in the distance estimates, and we may obtain better performance by estimating J for each time-frequency bin, as discussed in the next section.

4. EXPERIMENTAL RESULTS

Our first experiment, illustrated in Figure 2, consists of two microphones, five sources, and a line describing the virtual microphone path. In this example, each source consisted of a pure tone played at a different frequency. The interpolated audio was obtained using (20) where, rather than keeping α fixed, it ranged from -50 to 50, corresponding to a virtual microphone that slides from -5m to 5m over a period of 4 seconds. As expected, as the virtual microphone moves in front of each source position, that source becomes the loudest source present in the mixture. This effect is illustrated in Figure 3 by plotting the normalized power of each source as a function of the position of the virtual microphone. Figure 4 shows that the computation is accurate by comparing the estimated power for source 3 to the power computed analytically.

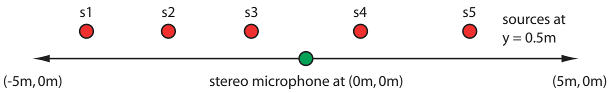

Figure 2: Virtual audio setup, experiment 1. Two microphones (M_0 and M_1) located at (-5 cm,0) and (5 cm, 0) capture the audio mixtures from five sources located at (-4.0 m, 0.5 m), (-2.5 m, 0.5 m), (-1.0 m, 0.5 m), (+1.0 m, 0.5 m), and (+3.0 m, 0.5 m). Each source emits a pure tune at a different frequency for 4 seconds, during which time a virtual microphone is moved at a constant speed from (-5,0) to (5,0).

Our second experiment had the same configuration as in Figure 2 with microphone separation 1 cm. Only the sources at positions $s1$ and $s5$ were active. The source at $s1$ was a female voice recording; the source at $s5$ was a male voice recording. As before, a virtual microphone was moved from -5m to 5m, this time over a period of 8 seconds. Figure 5 shows the result of applying (20) as in the first experiment to generate the interpolated audio signal. The dotted line in Figure 5 shows the theoretical signal-to-interference ratio (SIR) of the two sources at the virtual microphone. A positive SIR corresponds to dominance of the source $s1$, while a negative SIR cor-

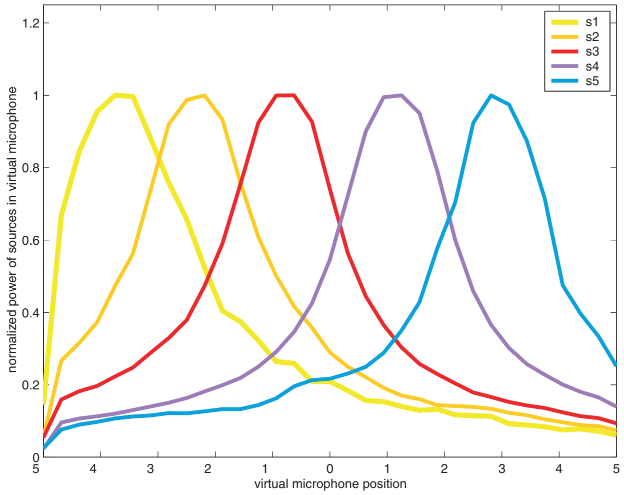

Figure 3: As the virtual microphone moves in front of each source, the relative power of that source reaches a maximum.

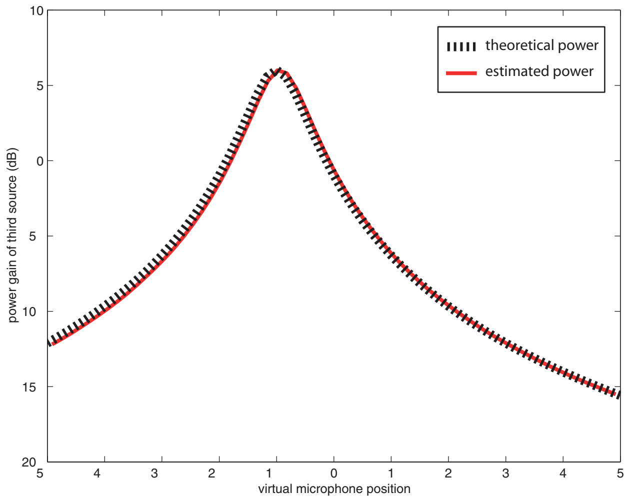

Figure 4: The dotted line is the theoretical power level of the third source at the moving the virtual microphone; the solid line is the power level produced by our audio interpolation method.

responds to dominance of the source $s5$. The solid line in Figure 5 is the instantaneous SIR of the interpolated mixture. The resulting relative strength of the voices in the virtual audio does not match the theoretical prediction as it did in the first experiment. Analysis of the distance estimation process revealed that the effect of the of the violations of the W-disjoint orthogonality assumption in the speech case caused errors in the $(\rho(\omega,\tau),\delta(\omega,\tau))$ estimates. The inaccurate estimates generated incorrect distance estimates that prevented the method from functioning as desired. In order to combat this, a preprocessing step was added to the $(\rho(\omega,\tau),\delta(\omega,\tau))$ estimation step. After the $(\rho(\omega,\tau),\delta(\omega,\tau))$ estimates were computed, a two-dimensional weighted histogram was generated to estimate the number of sources and their associated mixing parameters as in [5]. Each $(\rho(\omega,\tau),\delta(\omega,\tau))$ was then mapped to the closest estimated source mixing parameter pair yielding new estimates $(\rho'(\omega,\tau),\delta'(\omega,\tau))$ given by

$$\operatorname*{argmin}_{\substack{(\hat\rho_j,\hat\delta_j)\\ j=1,\ldots,\hat N}} |\log\rho(\omega,\tau)-\log\hat\rho_j|^2+|\delta(w,t)-\hat\delta_j|^2 \tag{22}$$

where $\hat N$ is the estimated number of sources and $(\hat\rho_j,\hat\delta_j)$, $j=1,\ldots,\hat N$, are the associated estimated relative amplitude and delay mixing parameters. Using these modified estimates, Figure 6 was generated. While the two curves still do not match exactly due to deviations from modeling assumptions, the interpolated audio curve has the correct character and the audio itself sounds realistic. The mixtures and generated virtual microphone sound files for both experiments are available online [9].

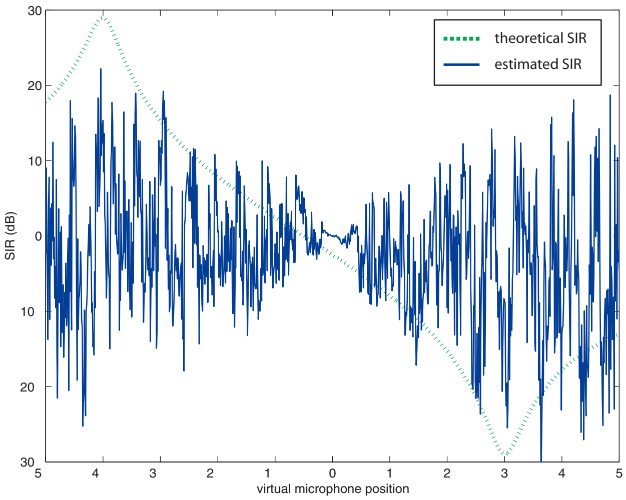

Figure 5: Experiment 2. SIR for theoretical (dotted) and interpolated (solid) mixtures, without explicit demixing.

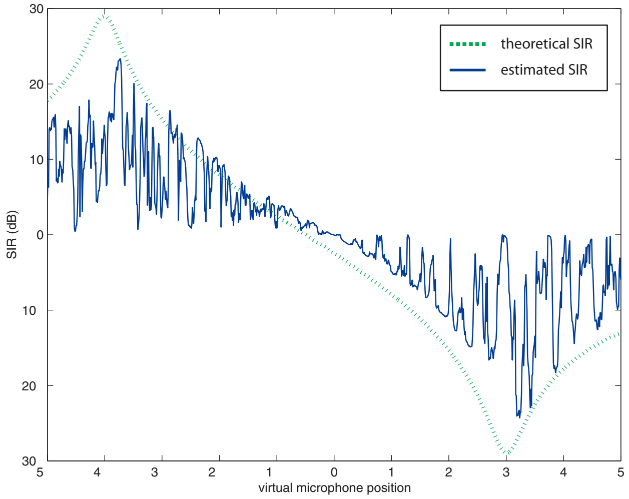

Figure 6: Experiment 2. SIR for theoretical (dotted) and interpolated (solid) mixtures, after explicit demixing.

5. SUMMARY

We presented a novel method for using a pair of audio mixtures to synthesize an accurate audio mixture from a different location. We emphasize that the method is blind and is not given any information about the number or location of the sources.

We note that in the case where the microphones and sources are coplanar, we could place a virtual microphone or trace a virtual microphone path anywhere in the plane, since (14)-(15) locate each source uniquely. We would need to alter (16) accordingly to represent the position of the virtual microphone with respect to the original pair.

Since the microphone separation needs to be small to resolve high frequencies, our algorithm is well suited to videoconferencing applications where the conversations of a roomful of people are captured by a single stereo microphone. Our method would allow the reconstruction of the conversations from a new location in the room using no additional information. However, we note that numerical considerations and deviations from the W-disjoint orthogonality assumption make the small-microphone- separation voice problem difficult, and we are currently working to address these issues.

REFERENCES

[1] S.E. Chen and L. Williams. View interpolation for image synthesis. In *Computer Graphics (SIGGRAPH '93)*, pages 279–288, July 1993.

[2] S.M. Seitz and C.R. Dyer. View morphing. In *Computer Graphics (SIGGRAPH '96)*, pages 21–30, August 1996.

[3] R. Radke, P. Ramadge, S. Kulkarni, T. Echigo, and S. Iisaku. Recursive propagation of correspondences with applications to the creation of virtual video. In *Proc. ICIP 2000*, September 2000. Vancouver, Canada.

[4] A. Jourjine, S. Rickard, and Ö. Yılmaz. Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures. In *Proc. ICASSP 2000*, volume 5, pages 2985–8, Istanbul, Turkey, June 2000.

[5] S. Rickard, Ö. Yılmaz, and A. Jourjine. On the approximate W-disjoint orthogonality of speech. In *Proc. ICASSP 2002*, Orlando, Florida, May 2002.

[6] M. Slaney, M. Covell, and B. Lassiter. Automatic audio morphing. In *Proc. ICASSP 1996*, May 1996.

[7] R. Balan, J. Rosca, S. Rickard, and J. O'Ruanaidh. The influence of windowing on time delay estimates. In *Proc. CISS 2000*, volume 1, pages WP1–(15–17), Princeton, NJ, March 2000.

[8] I. Daubechies. *Ten Lectures on Wavelets*, chapter 3. SIAM, Philadelphia, PA, 1992.

[9] R. Radke and S. Rickard. Audio interpolation page. `http://www.ecse.rpi.edu/homepages/rjradke/pages/ainterp/ainterp.html`. This page should be opened in a web browser with the Flash 5 plugin.