PHASE AND LEVEL DIFFERENCE FUSION FOR ROBUST MULTICHANNEL SOURCE SEPARATION

Johannes Traa[1] *Minje Kim*[2] *Paris Smaragdis*[1,2,3]

[1] University of Illinois at Urbana-Champaign, Department of Electrical and Computer Engineering
[2] University of Illinois at Urbana-Champaign, Department of Computer Science
[3] Adobe Systems Inc.

ABSTRACT

Inter-channel phase (IPD) and level (ILD) differences are common features in multichannel source separation algorithms like DUET and MENUET. However, their utility depends strongly on the configuration of the array and what microphone pairs are used to calculate them. IPDs are most useful when extracted from microphones that are close together as this avoids spatial aliasing. In contrast, ILD clusters are only well separated for widely spaced microphones. We investigate this trade-off between IPD and ILD features and propose a method to best combine them for multichannel source separation. Experimental results demonstrate the utility of this approach.

Index Terms— source separation, circular statistics, RANSAC

1. INTRODUCTION

Inter-channel phase (IPD) and level (ILD) differences have been used in source separation algorithms such as the Degenerate Unmixing Estimation Technique (DUET) [1] and its successor, Multiple sENsor dUET (MENUET) [2]. Joint IPD-ILD features form clusters corresponding to the directional sources in an audio mixture. Thus, the sources can be separated by applying the k-means algorithm (for example) to the set of extracted features and using the cluster labels to generate a binary mask for each source.

This approach works remarkably well in anechoic conditions for compact arrays thanks to the disjointness of speech signals in the time-frequency plane. However, it breaks down in real-world conditions due to reverberation and spatial aliasing. The goal of this paper is to develop a method for combining IPD and ILD features that maximizes separation performance in a more general, non-ideal setting.

In anechoic conditions, IPD features that correspond to a directional source are a linear function of frequency. However, if the microphones are further apart than a critical distance, spatial aliasing occurs. This causes the linear IPD function to wrap in $[-\pi, \pi]$. Aliasing has been addressed, for example, by modeling the features with directional probability densities such as the wrapped Gaussian [3], wrapped Laplacian [4], and von Mises [5], [6], [7] distributions.

In this paper, we discuss the impact of microphone spacing on the utility of both inter-channel features. In particular, we note that IPDs are most useful for separation when the microphones are relatively close together (e.g. < 10 cm). In this regime, IPDs form wrapped lines that can be identified even in reverberant, noisy conditions. However, ILDs are difficult to cluster because there is typically little difference in the signal amplitude between the channels. ILDs are most effective for separation purposes when the microphones are far apart (e.g. > 30 cm) [8]. But in this regime, IPDs will suffer from severe spatial aliasing. Thus, there is a crucial tradeoff in these features that depends on the configuration of the array.

It is interesting to note that this same trade-off is observed with binaural localization cues in human hearing [9, Chapter 13]. At lower frequencies, i.e. < 1 kHz, phase cues are most useful, at higher frequencies, attenuation cues are most useful, and in the intermediate regime, both are effective. This makes sense when we observe that the distance between a human's ears is roughly equal to half the wavelength of a 1 kHz sinusoid. Above this frequency, spatial aliasing takes its tole on phase cues and we must rely on attenuation cues instead.

In [7], the authors proposed an efficient method for clustering wrapped IPD features using the Random Sample Consensus (RANSAC) [10] algorithm. This approach reduced the problem of fitting multiple wrapped lines to IPD data to a simple sample-and-count procedure. We extend this approach to use both IPD and ILD features in combination with an arbitrary stereo microphone array.

2. INTER-CHANNEL FEATURES

In this section, we define inter-channel phase and level difference features as a function of the room impulse responses from a point source to a pair of microphones. We then discuss the impact of the array configuration on the utility of these features for blind separation. Finally, we introduce a probabilistic model for IPDs and ILDs.

2.1. Phase and level differences

Acoustic signals are recorded at either of two microphones. After applying the short-time Fourier transform (STFT) with window size $2D$, we have complex-valued matrices $\mathbf{X}^{(i)}, i = 1, 2$. The DFT coefficient at time frame t and frequency f is denoted as $X_{f,t}^{(i)}$. We will retain only the first D coefficients in any frame because the second half contains the same information. The element-wise log-ratio of the STFT matrices provides us with complex-valued inter-channel features for all time-frequency pairs of the form:

$$F_{f,t} = \log \left(\frac{X_{f,t}^{(1)}}{X_{f,t}^{(2)}} \right) \quad . \tag{1}$$

Phase differences can be defined as:

$$\delta_{f,t} = -\operatorname{Im}\left(F_{f,t}\right) = \angle X_{f,t}^{(2)} - \angle X_{f,t}^{(1)} \quad , \tag{2}$$

and level differences can be defined as:

$$\alpha_{f,t} = \operatorname{Re}\left(F_{f,t}\right) = \log\left(\left|X_{f,t}^{(1)}\right|\right) - \log\left(\left|X_{f,t}^{(2)}\right|\right) \quad . \tag{3}$$

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

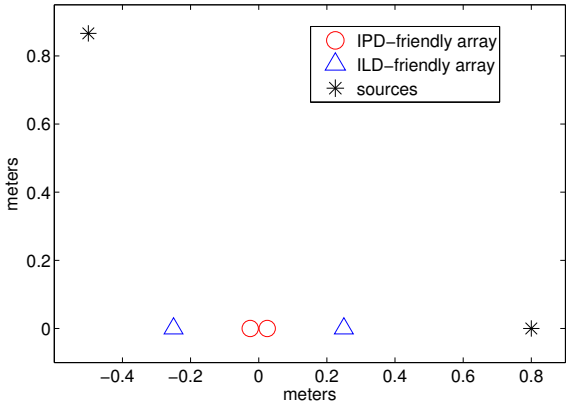

Fig. 1. Array configurations most amenable to either inter-channel feature. The microphones in the red (circle) and blue (triangle) arrays are separated by 5 and 50 cm, respectively.

One can show that, under anechoic conditions, the IPD and ILD features for a point source take the form:

$$\delta_{f,t} = \psi\left(\omega\left(d_1 - d_2\right)\right) \quad , \quad \omega = \frac{\pi f}{D} \quad , \tag{4}$$

$$\alpha_{f,t} = \log\left(a_1\right) - \log\left(a_2\right) \quad , \tag{5}$$

where ω is radian frequency, a_i and d_i are the attenuation and time delay due to propagation from the source to the i^{th} microphone, the wrapping function $\psi : \mathbb{R} \to \mathbb{S}$ is defined as:

$$\psi\left(x\right) = \text{mod}\left(x + \pi, 2\pi\right) - \pi \quad , \tag{6}$$

and $\mathbb{S} = \{\theta : \theta \in [-\pi, \pi]\}$.

We can see that inter-channel features are a function only of the room impulse responses from the source to the microphones. Thus, they depend on the physical configurations of the sources and of the array.

2.2. Impact of array configuration

It is worthwhile to carefully study the form of the features in (4)-(5). Fig. 1 depicts two stereo array configurations of interest along with two source positions. If the microphones are spaced far apart (blue triangles in Fig. 1), signals arriving along the axis of the array will induce a long delay $d_1 - d_2$ between the channels. This, in turn, gives rise to severely-wrapped IPD features. However, the difference in the log attenuations is sufficient to provide salient ILD features. IPDs and ILDs for this array configuration are shown in blue in the top panel of Fig. 2.

If the microphones are closely spaced (red circles in Fig. 1), IPD features are more useful as spatial aliasing has a limited effect. Unfortunately, the difference in attenuations $\log\left(a_1\right) - \log\left(a_2\right)$ is negligible, rendering ILD features mostly indistinguishable for either source. This is depicted in red in the bottom panel of Fig. 2.

Thus, there is a trade-off between the two features that must be taken into account when using DUET-style source separation algorithms with any given array. If 3 channels are available, we might place two close together and a third at a distance. This would make it possible to extract salient IPD features from the close pair and ILD features from one of these and the third microphone. However, this is a more expensive and constraining solution than the common 2-channel arrangement found, for example, in smartphones. In this paper, we will focus on how best to combine the features for a 2-microphone array.

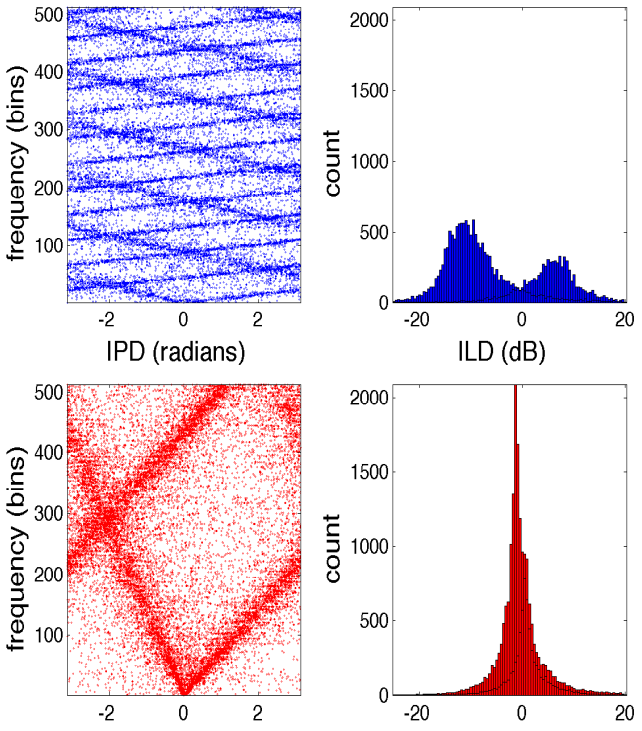

Fig. 2. IPD and ILD features associated with the arrays in Fig. 1. (Top row) ILD-friendly array. (Bottom row) IPD-friendly array.

2.3. Probabilistic model for IPDs and ILDs

When $K > 1$ speakers are active simultaneously, the features will form clusters corresponding to each one. This is due to the near-disjointness of speech in the STFT domain:

$$\forall\, f, t \quad \prod_{k=1}^{K} S_{f,t}^{(k)} \approx 0 \quad , \tag{7}$$

where $S^{(k)}$ is the STFT matrix of the k^{th} speech signal. Time-frequency overlap between sources, reverberation, and other interferences tend to smear the data and introduce outliers. This has the effect of broadening the clusters and increasing their overlap. To robustly perform clustering, we model either feature with an appropriate probability density.

We will model the IPDs, conditioned on the k^{th} source, with the von Mises (vM) [11] distribution:

$$v\mathcal{M}\left(\delta_{f,t}\,;\, h_k, \kappa\right) = \frac{1}{2\pi I_0\left(\kappa\right)} e^{\kappa \cos\left(\delta_{f,t} - h_k f\right)} \quad , \tag{8}$$

where h_k is the IPD line slope for the k^{th} source, κ is a concentration parameter, and $I_0\left(\kappa\right)$ is the 0^{th}-order modified Bessel function of the first kind. The conditional distribution of the ILDs is modeled as Gaussian:

$$\mathcal{N}\left(\alpha_{f,t}\,;\, \mu_k, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(\alpha_{f,t} - \mu_k\right)^2}{2\sigma^2}} \quad , \tag{9}$$

where μ_k is the mean for the k^{th} source and σ^2 depends on the spread of the ILD features.

3. CLUSTERING ALGORITHM

Let $\mathbf{Y}$ denote a dataset of N joint IPD-ILD features $\mathbf{Y}_n = \{\delta_n, \alpha_n\}$, $n = 1, \ldots, N$, extracted from a pair of microphones. To cluster these features, we define a likelihood function over the data set:

$$\mathcal{L}\left(\mathbf{Y}\,;\,\mathbf{h}, \boldsymbol{\mu}, \kappa, \sigma^2\right) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[v\mathcal{M}\left(\delta_n\,;\,h_k,\kappa\right)\mathcal{N}\left(\alpha_n\,;\,\mu_k,\sigma^2\right)\right]^{[z_n=k]} \quad , \qquad (10)$$

where $[-]$ denotes the indicator operator and z_n is the class label for the n^{th} time-frequency bin. Thus, we wish to maximize (10) with respect to the parameters:

$$\{\mathbf{h}, \widehat{\boldsymbol{\mu}, \kappa}, \sigma^2\} = \underset{\mathbf{h},\boldsymbol{\mu},\kappa,\sigma^2}{\operatorname{argmax}} \;\; \mathcal{L}\left(\mathbf{Y}\,;\,\mathbf{h}, \boldsymbol{\mu}, \kappa, \sigma^2\right) \quad . \qquad (11)$$

However, because the labels z_n are unknown and the IPD features form wrapped lines, this optimization is non-trivial. Instead of using an optimization routine that is susceptible to local optima, we will find the parameters with a variant of the RANdom SAmple Consensus (RANSAC) algorithm.

3.1. RANSAC

A RANSAC [10] algorithm was proposed in [7] for fitting a set of wrapped lines to an IPD dataset. We adapt it to the case of joint inter-channel features.

The RANSAC procedure to fit a single model is as follows. First, a set of M "samples" are selected uniformly at random from the dataset where each sample consists of the minimum number m of data points required to fit the model. For example, in the case of a line passing through the origin, a single point is required (i.e. $m = 1$). Each sample thus provides a candidate model. Inliers are counted for each candidate and the model with the highest count is chosen.

If the proportion of the dataset that consists of inliers is w, one can show that the expected number of RANSAC samples required to get a single inlier is w^{-m}. Typically, this is multiplied by a constant to ensure that a good model is fit (e.g. $M = 5\lceil w^{-m}\rceil$).

When $K > 1$ models are to be fit simultaneously, we can choose candidates in an iterative fashion. Alternatively, we may select K candidates in each RANSAC sample as proposed in the multiRANSAC algorithm [12]. We found that the latter strategy yielded superior results for joint IPD-ILD clustering.

We must define distance measures and thresholds to count inliers. To reflect the formulation in (10), we could choose the measures to be cosine and Euclidean distances and the thresholds to be $\tau_\delta = \sqrt{1/\kappa}$ and $\tau_\alpha = \sigma$ for the IPD and ILD features, respectively. We can now simply maximize an inlier count rather than (11). This works well in practice due to its efficiency and robustness to outliers.

3.2. Choice of thresholds

Due to the trade-off between IPD and ILD features discussed in Section 2.2, the choice of thresholds τ_δ and τ_α is crucial. When the microphones are far apart, we should set τ_δ to a large value and τ_α to a reasonably small value. This has the effect of ignoring the relatively uninformative IPD features. In contrast, when the microphones are close together, IPD features are more informative, and we should set the thresholds in the opposite manner.

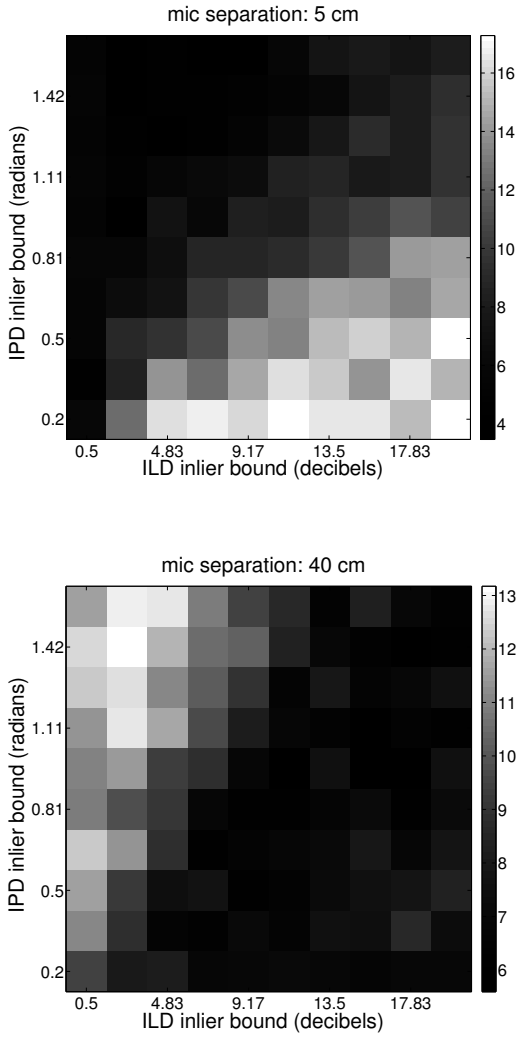

Fig. 3. Signal-to-Distortion Ratio averaged over 100 trials for closely- and widely-spaced microphone arrays.

We evaluated the Signal-to-Distortion Ratio (SDR) [13] over many pairs of thresholds to investigate what values would be best. The set-up for these experiments is as described in Section 5 and the results are shown in Fig. 3 with surface plots. The trade-off between the two features is evident. To ensure that our algorithm is robust to many possible array configurations, we randomly sample an inlier bound pair for each RANSAC sample. We found that, in practice, these can be selected uniformly at random from the union of two rectangular regions in τ_δ-τ_α space:

$$(\tau_\delta, \tau_\alpha) = \begin{cases} (\frac{\pi}{3}, 1.5) < (\tau_\delta, \tau_\alpha) < (\frac{\pi}{2}, 4.5) \\ (\frac{\pi}{16}, 15) < (\tau_\delta, \tau_\alpha) < (\frac{\pi}{8}, 20). \end{cases} , \qquad (12)$$

To prevent trivial solutions where a chosen pair of inlier bounds includes too many data points (e.g. ILD features are very similar for both sources but a wide inlier bound is chosen), we discard candidates with too many inliers. In practice, we found that candidates

that register more than 70% of the data as inliers should be ignored.

4. SOURCE SEPARATION

To perform source separation, we first estimate the model parameters $\mathbf{h}$ and $\boldsymbol{\mu}$. This is done with the sequential RANSAC procedure described in Section 3. Once the models are fit, we evaluate the posterior probabilities that each time-frequency data point belongs to each source. In accordance with (10), this is calculated as:

$$\eta_{nk} = \frac{v\mathcal{M}\left(\delta_n\,;\,\widehat{h}_k,\widehat{\kappa}\right)\mathcal{N}\left(\alpha_n\,;\,\widehat{\mu}_k,\widehat{\sigma}^2\right)}{\sum_{j=1}^{K} v\mathcal{M}\left(\delta_n\,;\,\widehat{h}_j,\widehat{\kappa}\right)\mathcal{N}\left(\alpha_n\,;\,\widehat{\mu}_j,\widehat{\sigma}^2\right)}\quad, \tag{13}$$

where the spread parameters $\widehat{\kappa}$ and $\widehat{\sigma}^2$ are determined by the optimal choice of inlier bounds from the simultaneous RANSAC results. Thus, we have $\widehat{\kappa} = \tau_\delta^{-2}$ and $\widehat{\sigma}^2 = \tau_\alpha^2$. The η_{nk}'s define soft time-frequency masks. To recover the k^{th} source, the mask with components $\boldsymbol{\eta}_k$ is multiplied element-wise with the mixture STFT from the first channel and the overlap-add algorithm is applied to reconstruct the time-domain waveform.

5. EXPERIMENTS

We evaluated the performance of our approach in a simulated reverberant room using the image method [14]. The room was of size 5×5 meters and the simulator was set up so that the T_{60} time was roughly 20 milliseconds.[1] The array was oriented horizontally and placed in the center of the room and the sources were located on the unit circle centered at the array. We conducted 100 trials in which two 2-3-second sentences were chosen at random from the TSP corpus [15] and mixed. The speakers were positioned in the middle of the room, separated by 2 meters, and the sensors were placed in-between the speakers with varying distances from 1 to 40 centimeters.

The sources were separated using masks derived via three methods: IPD only, ILD only, and IPD-ILD fusion (proposed). The IPD- and ILD-only model parameters were estimated as in the proposed method but using only one of the two features. Likewise, time-frequency masks were estimated using just the corresponding feature. 100 RANSAC samples were chosen for each trial.

The results from these experiments are summarized in Fig. 4. We can see that the proposed method is more robust to variability in the array configuration than either of the independent-feature methods. The trade-off discussed in Section 2.2 between IPD and ILD features as a function of microphone separation is also evident.

6. CONCLUSION

We have shown that there is a crucial trade-off between inter-channel level (ILD) and phase (IPD) difference features for audio source separation with a microphone array. When the microphones are closely spaced, IPDs provide more salient cues for distinguishing between the speakers. However, when the microphones are far apart, the opposite holds. When the array configuration is unknown or varying over time, separation quality may suffer from this trade-off. The proposed method uses a random sampling approach to leverage the strengths of both feature types over all array configurations.

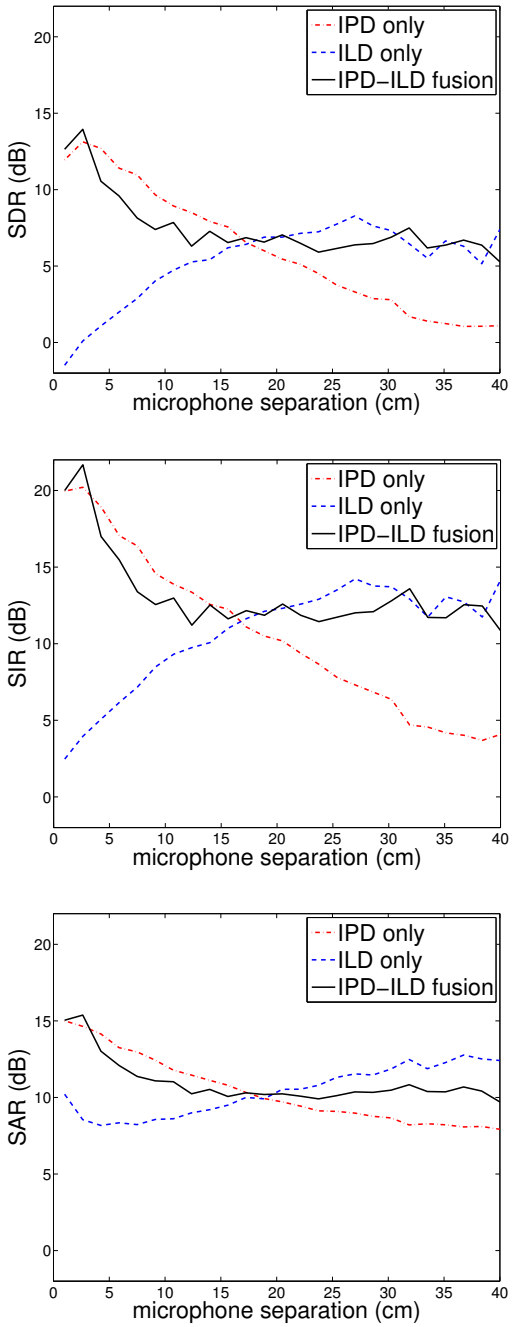

Fig. 4. Signal-to-Distortion, Signal-to-Interference, and Signal-to-Artifact Ratios for source separation experiments with three feature types.

[1] The T_{60} time of a room is how long it takes for the power of the room impulse response to drop by 60 dB.

7. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–1847, 2004.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833 – 1847, 2007.

[3] P. Smaragdis and P. Boufounos, "Learning source trajectories using wrapped-phase hidden Markov models," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 114–117, 2005.

[4] N. Mitianoudis, "A generalized directional Laplacian distribution: Estimation, mixture models and audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2397–2408, 2012.

[5] T. D. Downs and K. V. Mardia, "Circular regression," *Biometrika*, vol. 89, no. 3, pp. 683–697, 2002.

[6] C. Kim, C. Khawand, and R. M. Stern, "Two-microphone source separation algorithm based on statistical modeling of angular distributions," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4629–4632, 2012.

[7] J. Traa and P. Smaragdis, "Blind multi-channel source separation by circular-linear statistical modeling of phase differences," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[8] M. Kim, P. Smaragdis, G. G. Ko, and R. A. Rutenbar, "Stereophonic spectrogram segmentation using Markov random fields," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[9] A. H. Benade, *Fundamentals of Musical Acoustics*, Dover Publications, Inc, second edition, 1990.

[10] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[11] K. V. Mardia, "Statistics of directional data (with discussion)," *J. R. Statist. Soc.*, vol. B 37, pp. 349–393, 1975.

[12] M. Zuliani, C. S. Kenney, and B. S. Manjunath, "The multi-RANSAC algorithm and its application to detect planar homographies," in *IEEE International Conference on Image Processing*, 2005, vol. 3, pp. 153–6.

[13] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, 2006.

[14] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[15] Peter Kabal, "TSP Speech Database," 2002, Telecommunications and Signal Processing Lab, McGill University.