BLIND MULTI-CHANNEL SOURCE SEPARATION BY CIRCULAR-LINEAR STATISTICAL MODELING OF PHASE DIFFERENCES

Johannes Traa

University of Illinois at Urbana-Champaign

Paris Smaragdis

University of Illinois at Urbana-Champaign
Adobe Systems Inc.

ABSTRACT

We address the problem of blind separation of speech signals with a microphone array. We demonstrate that a signal propagating towards the array at an angle corresponds to inter-channel phase difference (IPD) data that lies on a wrapped line (i.e helix) in a circular-linear domain. Thus, the problem reduces to that of fitting helices to data that lies on a cylinder. However, outliers abound because of reverberation, noise, and signal overlap in the time-frequency domain, so we perform the clustering with a sequential variant of Random Sample Consensus (RANSAC). We show that this method can easily be applied to arrays with many microphones and that it is robust in reverberant experimental conditions.

***Index Terms*—** blind source separation, circular statistics, von Mises distribution, RANSAC

1. INTRODUCTION

The goal of this paper is to separate speech signals with an array of microphones in reverberant conditions. Various methods exist to tackle this problem. Beamforming [1] aims to amplify energy arriving from one or more directions while suppressing energy arriving at others. In contrast, time-frequency (TF) masking techniques rely on the sparsity of speech in the TF domain to partition audio mixtures into disjoint sets that correspond to the source signals. The Degenerate Unmixing Estimation Technique (DUET) [2] is one such approach that was originally applied to the case of 2-channel source separation. MENUET (Multiple sENsor dUET) [3] extends DUET to arrays with more than 2 microphones.

Although these methods are appealing, they fail to handle the issue of spatial/phase aliasing. To address this, the authors in [4] approximate inter-microphone delays by estimating the derivative of phase with respect to frequency. Alternative methods based on wrapped probability distributions explicitly model the phase differences between microphone pairs as circular variables. The Modified Discrete Cosine Transform (MDCT) was used in [5] to map the audio signals to a space where a mixture of wrapped Laplacian distributions can describe the audio mixture. Finally, in [6], the authors modeled phase differences of two signals with a two-component mixture of von Mises distributions.

In this paper, we model the phase differences in each time-frequency bin of the short-time Fourier transform (STFT) with the von Mises distribution. In addition, we take advantage of the linear relationship between phase difference and frequency to cluster the data by fitting multiple wrapped lines (one for each source). We then construct time-frequency masks based on this clustering to perform the separation.

The model-fitting step is challenging because phase information is very noisy in real-world environments and cross-over often occurs between the data corresponding to different sources. For these reasons, local optimization methods such as those in [3], [5] and [6] may fail to converge to a meaningful solution. We instead apply a sequential variant of Random Sample Consensus (RANSAC) to cluster the data. This method is robust in reverberant experimental conditions and can easily be used with many microphones.

The rest of this paper is organized as follows. In section 2, we discuss the circular-linear model for inter-channel phase difference data and cast the source separation problem as one of circular-linear regression. In section 3, we describe a sequential RANSAC algorithm and detail its application to source separation. In section 4, we discuss how a phase-wrapped model is beneficial and in section 5, we show experimental results. Section 6 concludes the paper.

2. CIRCULAR-LINEAR MODEL

2.1. IPDs as circular-linear data

We will use inter-channel phase differences (IPD) to distinguish between speakers. It was shown in [2] that blind source separation is possible using just IPDs and inter-channel level differences (ILD) with DUET. The primary assumption is that the source signals are disjoint in a time-frequency representation, i.e. the energy in each bin of the mixture's STFT is dominated by one source. If this condition holds, the mixture STFT can be partitioned such that only the bins assigned to the j^{th} source are used to reconstruct it (time-frequency masking). In practice, speech signals are nearly disjoint.

In DUET, signals arriving at distinct angles are separated

with a two-microphone array. The IPD information is calculated as the element-wise differences between the phase components of the two channel STFTs $X^{(i)}, i = 1, 2$. These differences are then normalized by frequency and collected in a histogram in the range $[-\pi, \pi]$. Peaks will occur at IPDs corresponding to delays between the two microphones. Unfortunately, delays of more than one sample cause phase-wrapping that corrupts the IPD histogram. To remedy this, we omit normalization and construction of the histogram and explicitly model phase differences as circular variables:

$$\delta = \angle X^{(1)} - \angle X^{(2)} \ . \tag{1}$$

A signal that arrives at an angle incurs a delay between the microphones. By the delay property of the Fourier transform, this corresponds to a phase shift in the frequency domain. More shift will exist at higher frequencies, resulting in data that lies along a wrapped line in a plot of frequency against IPD. An example of this for a synthetic, anechoic mixture of three sources is shown in Fig. 1(a) with the source lines superimposed. When we reshape this plot as in Fig. 1(b), we see that a wrapped line is essentially a helix in a cylindrical domain. Thus, we can perform source separation by fitting multiple helices to this data.

2.2. Circular-linear regression

From subsection 2.1, it is clear that the source separation problem reduces to one of circular-linear regression. We first look at the case of fitting a single wrapped line. Given N measurements of two variables $y_i = (\delta_i, f_i)$, $i = 1, \ldots, N$, we would like to fit a wrapped line (a.k.a. barber-pole regression curve [7]) of the form $\delta = \text{mod}(\alpha f , 2\pi)$ to this data. In this paper, δ and f represent IPD and frequency, respectively, and each y_i corresponds to a single time-frequency bin.

We use the von Mises distribution [8] to measure error in the circular variable δ. Its probability density function, parameterized by mean μ and concentration k, has the form

$$P(x|\mu, k) \propto e^{k \cos(x-\mu)} \ . \tag{2}$$

Now we can describe the helix-fitting problem. We wish to find the slope $\hat{\alpha}$ that maximizes the sum of von Mises log likelihoods in (3). Maximizing in (3) is equivalent to minimizing a sum of distances in δ from the wrapped line to the data. Note that $\mu_i = \alpha f_i$ depends on the data index i.

$$\hat{\alpha} = \underset{\alpha}{\text{argmax}} \prod_{i=1}^{N} P(y_i|\mu_i, k) = \underset{\alpha}{\text{argmax}} \sum_{i=1}^{N} \cos(\delta_i - \alpha f_i) \tag{3}$$

Unfortunately, local optimization is unreliable here because of the many local maxima caused by outliers and wrapping in the circular variable. This is especially true when fitting multiple lines. It may also be prohibitively expensive in real-time applications with large arrays. We might be tempted to scan over the space of all physically possible delays, but this is unnecessary. Instead, we will use a data-driven search strategy with random sampling.

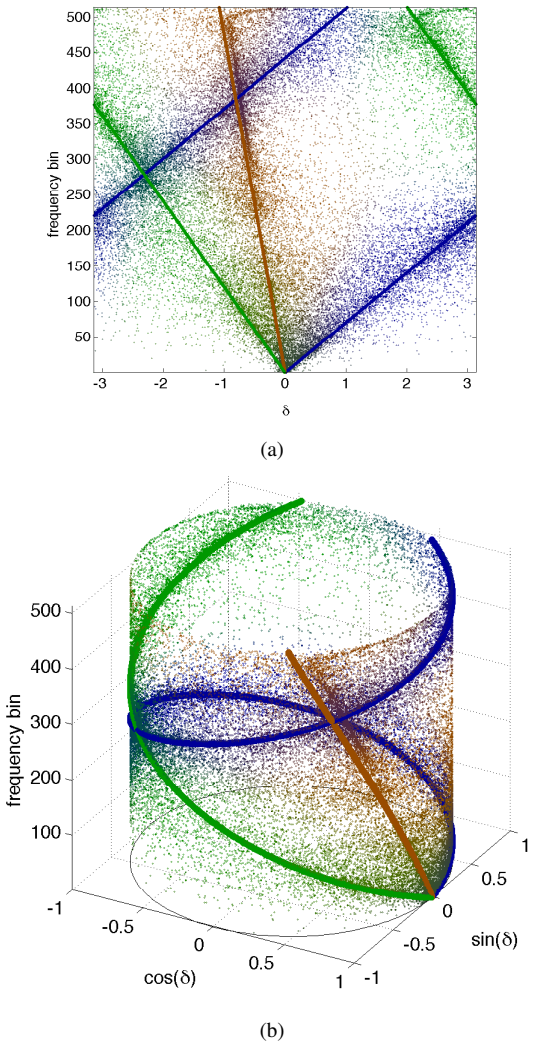

Fig. 1. (a) IPD plot for synthetic mixture of three sources, colored according to likelihood probability (see subsection 2.2). (b) Equivalent cylinder visualization.

3. CLUSTERING AND SOURCE SEPARATION

3.1. Sequential Random Sample Consensus

We now describe our method for clustering the IPD data. Random Sample Consensus (RANSAC) [9] is known in the computer vision community to be a robust method for estimating simple models such as lines and circles in the presence of outliers. The underlying principle is that the true model can be described well by a number of data points. For a line

Algorithm 1 Sequential RANSAC

Inputs: $x : N$ data points
K : number of models to fit
Output: $\hat{\theta} : K$ source models

$y = M$ samples from x selected uniformly at random
$I = 0^{N \times M}$
for $i = 1 : M$ **do**
  Fit model θ_i to y_i
  $I(n, i) = 1$, $\forall n$ s.t. x_n is inlier of θ_i
$\hat{\theta} = \{\}$
$A = \{1, \ldots, N\}$
for $j = 1 : K$ **do**
  $\hat{i} = \underset{i}{\operatorname{argmax}} \sum_{n \in A} I(n, i)$
  $\hat{\theta} = \hat{\theta} \cup \theta_{\hat{i}}$
  $A = A \setminus \{n : I(n, \hat{i}) = 1\}$
return $\hat{\theta}$

passing through the origin, which is the case we address in this paper, we only need one point to fully specify the model. To select the correct data for fitting, candidates (samples) are chosen at random such that at least one of them coincides with the true model with high probability.

We can ensure that a good model is fit with high probability by selecting enough samples. The number M of samples is determined by the expected number of trials $\mathrm{E}[t]$ until an inlier is chosen. If the probability of choosing an inlier is p, it can be shown ([9]) that $\mathrm{E}[t] = p^{-1}$. In practice, we set $M = C\ \mathrm{E}[t]$, with $C > 1$ to ensure that an inlier is sampled. In our experiments, we found that $C = 10$ was sufficient.

Fast, sequential variants of RANSAC have been proposed to identify multiple planar homographies for a stereo imaging application in [10] and [11]. We apply a similar approach to perform multi-model, circular-linear regression (Algorithm 1). The process requires that M be scaled proportionally with the number of sources K and that once a source model is chosen, all inliers be removed. Note that this algorithm makes no assumptions about the form of the model θ and that, in our particular application, θ takes the form of a wrapped line.

3.2. Blind source separation

To perform blind source separation, we extract IPD features, cluster them, and construct probabilistic masks to reconstruct the individual sources. We first discuss stereo unmixing and then generalize to the case of two or more channels.

Phase differences between the two channels are calculated as in (1), resulting in N phase difference-frequency pairs $y_i = (\delta_i, f_i)$. Sequential RANSAC can now be applied to fit K helices with slopes α_j, as in Fig. 1. In our experiments, the i^{th} data point is considered an inlier of the j^{th} helix if δ_i is within $\pm\frac{\pi}{8}$ of the helix value $\mu_{ij} = \alpha_j f_i$ (modulo 2π).

To recover the K source signals, we apply time-frequency masks to one of the mixture STFTs. To calculate the masks, probabilities are evaluated for every data point (i.e. time-frequency bin) according to the von Mises likelihood criterion

$$\forall\, i, j \quad w_{ij} = \frac{P(y_i|\mu_{ij}, k)}{\sum_{l=1}^{K} P(y_i|\mu_{il}, k)} = \frac{e^{k\cos(\delta_i - \alpha_j f_i)}}{\sum_{l=1}^{K} e^{k\cos(\delta_i - \alpha_l f_i)}}\ , \quad (4)$$

where $\mu_{ij} = \alpha_j f_i$ depends on both the data and helix indices.

These probabilities represent how likely it is that the i^{th} bin belongs to the j^{th} source. To reconstruct source j, (4) is multiplied by the corresponding bins in the mixture STFT and the result is transformed to the time domain with the inverse STFT. We can achieve more aggressive separation by increasing the concentration parameter k. In the limit as $k \to \infty$, (4) reduces to a maximum-likelihood binary mask where each bin contributes to the reconstruction of a single source:

$$\forall\, i \quad w_{ij}^{b} = \begin{cases} 1 & \text{if} \quad w_{ij} = \max_{l}\ w_{il} \\ 0 & \text{else} \end{cases}\ . \quad (5)$$

We can also apply this technique with more than two channels. In the case of three channels, two informative delays are present from microphone pairs 1-2 and 1-3. The corresponding IPD data has two circular axes instead of one. This can only increase the inter-cluster distances, which leads to better clustering. In this way, sequential RANSAC can easily be applied to arrays with two or more microphones.

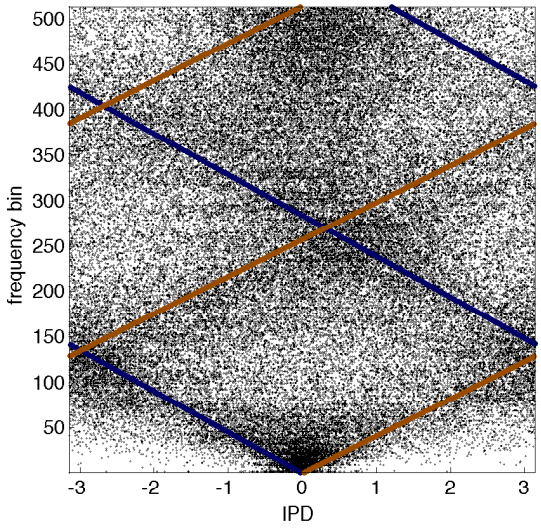

Fig. 2. IPD plot of highly reverberant ($T_{60} = 1.5$ seconds), 2-speaker stairwell recording. Sequential RANSAC succeeds in identifying the source models in the presence of 65% outliers.

4. BENEFITS OF A PHASE-WRAPPED MODEL

There are two reasons why explicitly modeling wrapping in the phase differences is advantageous. First, the circular-

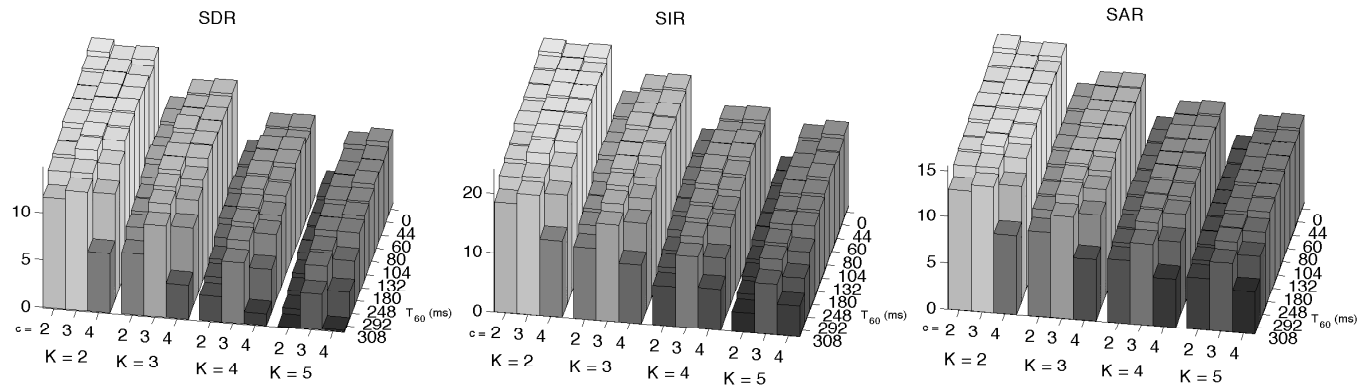

Fig. 3. Signal-to-distortion, signal-to-interference, and signal-to-artifact ratios for 2- and 3-channel source separation in a 2D room and 4-channel source separation in a 3D room. c is the number of channels and K is the number of sources.

linear model allows delays of more than one sample. This is useful in the physical world because microphone pairs need not be very close to one another. The maximum allowed distance can be expressed as

$$d = \frac{mv}{s}\ , \tag{6}$$

where m is the maximum expected delay, v is the speed of sound, and s is the sampling rate. Note that $2m$ is the most full wrap-arounds allowed in the IPD plot.

As shown in [2], the maximum inter-microphone distance when $s = 16000$, $v = 340$, and $m = 1$ is $d = 2.125$ cm. Our method delivers reliable estimates for up to 6 samples of delay, for which the distance could increase six-fold: $d = 12.75$ cm. This can be used for leveraging attenuation information in reverberant environments that is not present with closely-spaced microphones. Conversely, we can separate high-quality audio with a sampling rate of 48 kHz using an array with $d = 4.25$ cm.

A second advantage of wrapping is the increased separation of the data. Although some cross-over occurs in the IPD plot, an increased separation overall improves the clustering (especially in the low frequencies).

5. EXPERIMENTS

To test our method, we simulated a reverberant room with an array of omnidirectional microphones placed at its centroid. We mixed two-second audio clips of five speakers from the TSP corpus [12] with speakers located at random but distinct angles on the unit semicircle, unit circle, and unit sphere for 2-, 3-, and 4-channel unmixing, respectively. The microphones were positioned 5 cm apart in a right-angle configuration. We ran 100 trials for $K = \{2, 3, 4, 5\}$ speakers with randomly-chosen sentences downsampled to 16 kHz. STFTs were calculated with a 1024-sample window and $\frac{3}{4}$-overlap and the overlap-add algorithm was used to invert them.

To simulate reverberation, we used the image method [13]. The T_{60} time (required for reverberant energy to drop 60 dB below direct path energy) of the room was varied from 0 to 308 milliseconds. We tested 2- and 3-channel separation in a 2D room (5×5 m) and 4-channel separation in a 3D room ($5 \times 5 \times 5$ m). We evaluated the performance of our method with the BSS Eval toolbox [14] using the individual sources convolved with the appropriate room impulse responses as the reference signals. This is so that we test for source separation rather than de-reverberation.

Fig. 3 summarizes the average performance of our method using a binary mask. These experiments show that increasing the number of microphones improves the separation by reducing overlap in the IPD data. However, the decrease in performance from 2D to 3D simulations suggests that early reflections have a noticeable impact on separation quality. Also, the audio outputs remain reverberant, so further processing is necessary to fully recover the original source signals.

To test our method's robustness to severe, real-world reverberation, the authors recorded simultaneous speech with a stereo iMic recorder in various indoor locations. The most difficult case was in a stairwell with a T_{60} time of 1.5 seconds. The IPD plot for this recording (Fig. 2) shows that the algorithm succeeded in identifying the correct source models even in such harsh conditions. Furthermore, a subjective assessment confirmed that the speakers were indeed separated.

Finally, we note that the time complexity of Algorithm 1 is roughly $\mathcal{O}\left(MN(3c + K + 1)\right)$, where M, N, c and K are the numbers of samples, time-frequency bins, channels and sources, respectively.

6. CONCLUSION

In this paper, we introduced a sequential RANSAC-based method for blind, multi-channel source separation. It improves on previous methods in three ways. First, the issue of wrapping in the inter-channel phase differences as discussed in [2] is mitigated by explicitly modeling phase as a circular variable. Second, source models are fit with a sampling technique that is robust to outliers, which are common in real-world audio data. And finally, the proposed algorithm is shown to be trivially applied to arrays with any number of microphones.

7. REFERENCES

[1] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Topics in Signal Processing: Microphone Array Signal Processing*, vol. 1, Springer, 2008.

[2] 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.

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

[4] Y. Wang, O. Yilmaz, and Z. Zhou, "Phase aliasing correction for robust blind source separation using duet," *IEEE Transactions on Signal Processing*, 2011.

[5] 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.

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

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

[9] 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.

[10] E. Vincent and R. Laganiere, "Detecting planar homographies in an image pair," *2nd International Symposium on Image and Signal Processing and Analysis*, pp. 182–187, 2001.

[11] Y. Kanazawa and H. Kawakami, "Detection of planar homographies with uncalibrated stereo using distribution of feature points," *British Machine Vision Conference*, vol. 1, pp. 247–256, 2004.

[12] Peter Kabal, "Tsp speech database," 2002, Telecommunications and Signal Processing Lab, McGill University.

[13] 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.

[14] 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.