Multichannel Source Separation and Tracking with RANSAC and Directional Statistics

Johannes Traa, *Student Member, IEEE,* and Paris Smaragdis, *Senior Member, IEEE*

***Abstract*—We describe multichannel blind source separation and tracking algorithms based on clustering wrapped inter-channel phase difference (IPD) features. We pose the clustering problem as one of multimodal circular-linear regression and present its probabilistic formulation. Phase wrapping due to spatial aliasing is explicitly incorporated by modeling the IPD features as circular variables. We present two methods based on Expectation-Maximization (EM) and a sequential variant of RANdom SAmple Consensus (RANSAC). We show that their strengths can be combined by using RANSAC to initialize EM.**

The IPD clustering algorithm is applied to separate stationary speakers from a multichannel mixture. We then extend it to the case of moving speakers by tracking their directions-of-arrival with the Factorial Wrapped Kalman Filter (FWKF) using RANSAC as a data pre-processor. Experimental results demonstrate that the proposed methods perform well in the presence of reverberant babble noise and spatial aliasing. The FWKF successfully tracks and separates moving speakers with separation quality comparable to that for stationary speakers.

***Index Terms*—directional statistics, interchannel phase difference, blind source separation, wrapped Kalman filter**

I. Introduction

In this paper, we are interested in separating (possibly moving) speakers who are in the far field region of a compact microphone array. Formally, multichannel Blind Source Separation (BSS) is the inverse problem of recovering K unknown source signals from C observed mixtures (one from each microphone). Beamforming methods [1] approach BSS from an array processing perspective. A spatial filter is designed so as to allow signals impinging on the array from particular directions to pass undistorted while blocking interfering signals incident at other angles. The Delay-and-Sum (DS) and Linearly-Constrained Minimum Variance (LCMV) beamformers are well-known examples. Many variants have been proposed including the Speech Distortion-Weighted Multichannel Wiener Filter (SDW-MWF) [2], which seeks to balance noise reduction with speech distortion.

Another famous approach known as Independent Components Analysis (ICA) [3], [4] attempts to invert a mixing matrix that relates the sources to the mixtures. A third approach to BSS, which we adopt in this paper, is the Degenerate Unmixing Estimation Technique (DUET) [5] and its extension to more than 2 sensors [6]. These algorithms cluster inter-channel phase and level differences (IPD, ILD) to construct a time-frequency (TF) mask and are known to produce extremely clean speech separation results in non-reverberant environments. Three assumptions made in DUET are that (1) the source signals are approximately disjoint in a TF representation, (2) at most one sample of delay is observed between the channels, and (3) early reflections are negligible. Speech signals are remarkably disjoint in the short-time Fourier transform (STFT) domain, even in the presence of strong reverberation [7]. Thus, the first assumption often holds. The second assumption is violated for high sampling rates or arrays with more than a few centimeters of separation between the microphones due to spatial aliasing. Solutions include oversampling [8] and explicit modeling of phase as a wrapped quantity [7], [9], [10]. We will adopt the latter approach in this paper. The third assumption is violated when there are strong early reflections from objects near the array [11].

These algorithms are often further extended to separate moving targets by tracking their positions over time. This can have a major impact on source separation performance. In the experiments section, we report improvements over the proposed batch algorithm of up to 7 dB by adapting it to track the sources. This requires that we estimate some quantity related to the source position such as time-delay-of-arrival (TDOA) or direction-of-arrival (DOA). The Generalized Cross Correlation (GCC) method [12], [13] computes pair-wise channel correlations on a short-term basis and looks for peaks in the resulting function over TDOA space. A generalization to more than two channels called Multi-Channel Cross Correlation (MCCC) was proposed in [14].

For compact arrays with 1-10 centimeters of spacing between the microphones, accurately estimating TDOAs can be difficult in a noisy environment. In this case, it is more appropriate to estimate the DOAs of the sources. The Steered Response Power (SRP) method scans DOA space with a beamformer and looks for peaks in the output power. A potential downside of this approach is that the SRP must be computed for each search direction on a grid. In addition, the resolution over DOA space is poor for arrays with closely-spaced elements. In this case, a more effective approach is the MUltiple SIgnal Classification (MUSIC) [15] algorithm, which identifies signal and noise subspaces of a channel correlation matrix. A "MUSIC spectrum" is calculated that contains peaks at the source DOAs. This requires that there be more channels than sources (i.e. $C > K$). To avoid having to scan over DOA space, search-free variants of MUSIC have been proposed such as root-MUSIC [16] and ESPRIT [17].

Tracking on the unit circle shows up in many contexts

J. Traa is a PhD student in the ECE department at the University of Illinois at Urbana-Champaign (UIUC) (traa2@illinois.edu).

P. Smaragdis holds a joint faculty position in ECE and CS at UIUC and works with Adobe Systems, Inc. (paris@illinois.edu)

including localization [1], phase-locked loops [18], and phase unwrapping [19]. We will represent the source locations by their azimuth angles (DOAs) and model these with the wrapped Gaussian (WG) distribution [20]. This distribution was used to learn source trajectories with a wrapped-phase HMM [21] and more recently to derive the Wrapped Kalman Filter (WKF) [22] and its multi-target variant, the Factorial WKF (FWKF) [23]. The Kalman filtering framework [24] is convenient in that it allows for sophisticated, statistically-grounded approaches to the tracking problem. A related method [25] uses DUET-style TF masking to follow multiple moving sources with a particle filter [26].

In previous work [7], we used the von Mises distribution [20] to model wrapped IPD features as circular-linear data [27], reducing the underdetermined BSS problem to one of multimodal circular-linear regression. We applied a sequential variant of the RANdom SAmple Consensus (RANSAC) algorithm [28] to perform the regression and cluster the features. We will explore the wrapped clustering problem in more detail and introduce an Expectation-Maximization (EM) [29] algorithm to solve it. We will then show that the RANSAC procedure can be used to initialize EM so as to avoid local optima and speed convergence. Finally, we apply RANSAC as a pre-processor in a FWKF to perform efficient, on-line tracking and separation.

The contributions of this paper are:

- a probabilistic formulation for the wrapped IPD clustering problem and an EM algorithm to solve it
- a detailed account of the RANSAC-based source separation algorithm presented in [7]
- an extension of this algorithm for non-stationary sources via the Factorial Wrapped Kalman Filter [22]
- experimental validation of the proposed methods

II. Directional Statistics

Directional statistics [20] is concerned with the analysis of quantities that lie on a circle, torus, or sphere. In this paper, we will find use for two directional distributions on the unit circle: the von Mises and wrapped Gaussian.

A. Unit circle

The unit circle is the 1D subspace of $\mathbb{R}^2$ consisting of all unit vectors $\mathbf{x}$:

$$\mathbb{S}^1 = \{\mathbf{x} : \|\mathbf{x}\|_2 = 1 \ , \ \mathbf{x} \in \mathbb{R}^2\} \ . \tag{1}$$

Alternatively, we can represent each unit vector with the scalar angle $\theta = \angle\mathbf{x}$. Thus, $\mathbb{S}^1$ can be interpreted as a 1D interval, where the boundaries represent the same value:

$$\mathbb{S}^1 = \{\theta : \theta \in [-\pi, \pi]\} \ . \tag{2}$$

We will find the latter representation more useful.

B. Wrapped distributions

We will model wrapped quantities as circular random variables $\theta \in \mathbb{S}^1$ with the von Mises (vM) and wrapped Gaussian (WG) distributions. We define a useful mapping $\psi : \mathbb{R}^1 \to \mathbb{S}^1$ that folds the real line around the unit circle:

$$\psi(x) = \text{mod}(x + \pi, 2\pi) - \pi \ . \tag{3}$$

1) Wrapped Gaussian (WG): The wrapped Gaussian arises from applying (3) to a Gaussian random variable on $\mathbb{R}^1$:

$$p\left(\theta \,;\, \mu, \sigma^2\right) = \sum_{l=-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-(\mu+2\pi l))^2}{2\sigma^2}} \ . \tag{4}$$

We can visualize it on the unit circle in $\mathbb{R}^2$ or directly in $\mathbb{S}^1$ (see Fig. 1). We will use the WG to cluster IPD features in an EM algorithm and to model the speaker DOAs in the FWKF.

2) von Mises (vM): The von Mises distribution is parameterized by a mean μ and a concentration κ:

$$p\left(\theta \,;\, \mu, \kappa\right) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta-\mu)} \ , \tag{5}$$

where $I_0(\kappa)$ is the 0^{th}-order modified Bessel function of the first kind. It is derived by conditioning a 2D Gaussian, $\mathcal{N}\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}\right)$, $\|\boldsymbol{\mu}\|_2 = 1$, on the unit circle and converting from Cartesian to polar coordinates [30]. The conditioning is such that $\kappa = 1/\sigma^2$. We choose the vM for its simplicity to model IPD features in the RANSAC clustering algorithm.

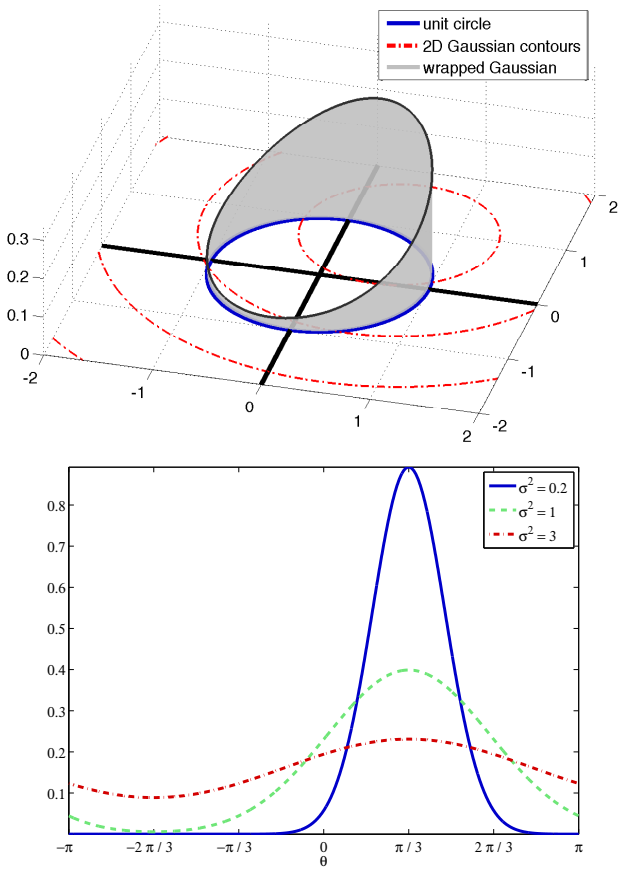


Fig. 1. (Top) Wrapped Gausian pdf ($\mu = \frac{\pi}{3}$) on the unit circle in $\mathbb{R}^2$ shown with 2D Gaussian contours ($\sigma^2 = 0.8$). (Bottom) WG pdf in $[-\pi, \pi]$ ($\mu = \frac{\pi}{3}$). The θ axis is the unit circle, unfolded. (This figure appears in [22].)

III. Inter-Channel Phase Difference Features

We will use inter-channel phase differences (IPD) as features to perform joint multi-source separation and tracking. To account for spatial aliasing, the IPD representation is modified from that of [5] so as to incorporate phase wrapping explicitly in a statistical model (this is similar to the approach taken in [31]). We show that these wrapped IPDs compose a circular-linear dataset and present a probabilistic interpretation of the regression problem.

A. Feature Extraction

A microphone array captures C time-domain signals that are converted to a time-frequency representation using the Short-Time Fourier Transform (STFT):

$$\mathbf{X}^{(i)} \in \mathbb{C}^{D \times T} \quad , \quad i = 1, \ldots, C \quad , \tag{6}$$

where D denotes the coefficient index in the DFT corresponding to half the sampling rate and T denotes the number of frames captured. We ignore the second half of the DFT because it contains the same information as the first half.

Since the Fourier transform is a linear operation, we have that the DFT coefficient at each time-frequency bin is equal to the sum of the contributions from the sources. In the absence of reverberation, this gives:

$$X_{ft}^{(i)} = \sum_{j=1}^{K} S_{ft}^{(j)} a_{ij} \, e^{-j\omega d_{ij}} \quad , \quad \omega = \frac{\pi f}{D} \quad , \tag{7}$$

where $S_{ft}^{(j)}$ is the DFT coefficient of the j^{th} source, a_{ij} and d_{ij} are the attenuation and delay for the direct path between the i^{th} microphone and the j^{th} source, and ω is the digital frequency corresponding to the f^{th} frequency band.

We compute element-wise logratios to consolidate the STFT information across pairs of channels. When, for example, $C = K = 2$, we have:

$$\begin{aligned} F_{ft} &= \log\left(\frac{X_{ft}^{(1)}}{X_{ft}^{(2)}}\right) \\ &= \log\left(\frac{S_{ft}^{(1)} a_{11} e^{-j\omega d_{11}} + S_{ft}^{(2)} a_{12} e^{-j\omega d_{12}}}{S_{ft}^{(1)} a_{21} e^{-j\omega d_{21}} + S_{ft}^{(2)} a_{22} e^{-j\omega d_{22}}}\right) \quad . \end{aligned} \tag{8}$$

If the signals are assumed to be approximately disjoint in the STFT domain [5], i.e.:

$$\forall\, f, t \quad S_{ft}^{(1)} S_{ft}^{(2)} \approx 0 \quad , \tag{9}$$

then we can simplify (8) to the one-source case:

$$F_{ft} \approx \log\left(\frac{S_{ft} a_1 e^{-j\omega d_1}}{S_{ft} a_2 e^{-j\omega d_2}}\right) = \log\left(\frac{a_1}{a_2}\right) - j\omega\,(d_1 - d_2) \quad . \tag{10}$$

The negative imaginary part of (10) yields the IPD:

$$\delta_{ft} \;=\; -\,\mathrm{Im}(F_{ft}) = \omega\,(d_1 - d_2) = \angle X_{ft}^{(2)} - \angle X_{ft}^{(1)} \quad . \tag{11}$$

Thus, the features lie on a wrapped line in a plot of frequency versus phase difference:

$$\delta_{ft} = \psi(\alpha\, f) \quad , \quad \alpha = \frac{\pi}{D}(d_1 - d_2) \quad , \tag{12}$$

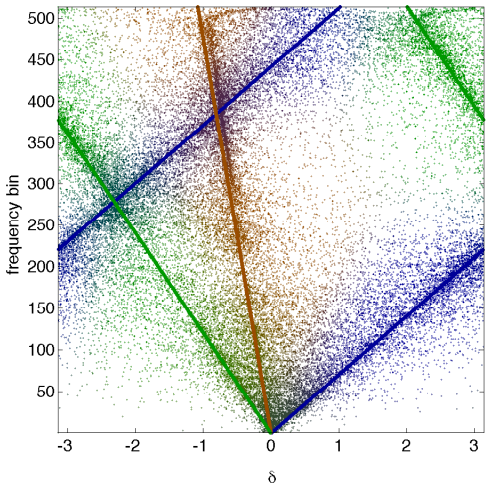


Fig. 2. IPD plot for a synthetic mixture of three speakers, colored according to von Mises assignment probability. The mean lines under this model are superimposed. (This figure appears in [7].)

where $\psi(-)$ is defined in (3). To make the dependence on frequency explicit, we form the following feature vector:

$$\boldsymbol{\delta}_{ft} = [\delta_{ft}\,,\, f] \quad . \tag{13}$$

Now it is clear that a collection of these vectors composes a circular-linear dataset. The case of three or more sources ($K \geq 3$) is no different as long as the disjointness property (9) holds for all source pairs. For three or more microphones ($C \geq 3$), the IPD feature vector contains $C - 1$ phase differences:

$$\boldsymbol{\delta}_{ft} = [\delta_{ft}(1,2), \ldots, \delta_{ft}(1,C)\,,\, f] \quad , \tag{14}$$

where $\delta_{ft}(1,i)$ denotes the IPD calculated from the 1^{st} and i^{th} channels. This representation is similar to that in [6].

B. IPD features as circular-linear data

We have seen that an acoustic wavefront that arrives at a microphone array at an angle induces a time delay between the microphones that 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.[1] When multiple speakers are present and they satisfy (9), we observe IPDs that trace out multiple wrapped lines. An example of this for a synthetic, anechoic mixture of three sources is shown in Fig. 2. We can only expect to locate the lines when the microphones are sufficiently closely-spaced since, otherwise, extreme wrapping effects arise. At a sampling rate of 16 kHz, a reasonable upper limit on inter-channel spacing is 10 cm.

To perform IPD-based BSS, we must cluster the features (14) and partition the mixture STFT accordingly. This is equivalent to performing multimodal circular-linear regression, namely, recovering the underlying wrapped linear models.

[1] This has also been called a "barber pole regression curve" in the directional statistics literature [27] because it can be visualized as a helix on the surface of a cylinder.

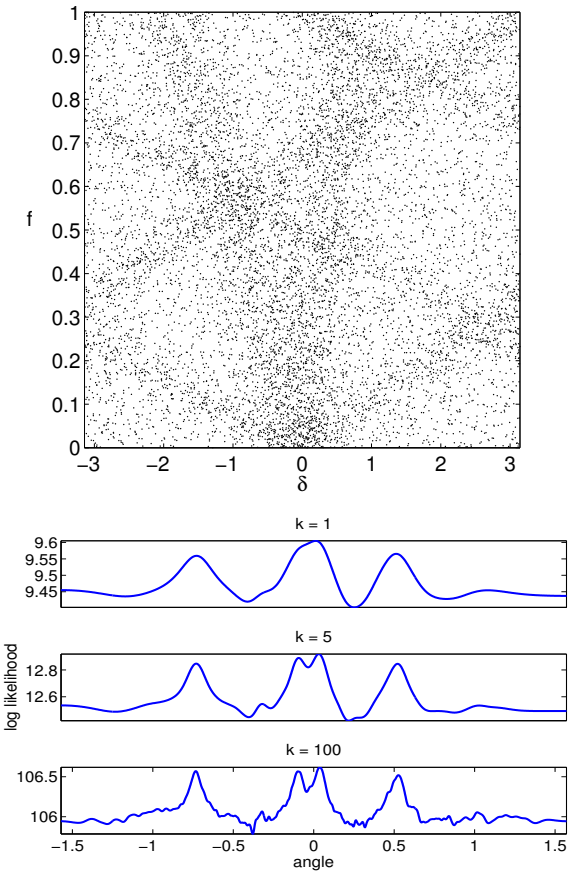


Fig. 3. (Top) 10,000 circular-linear data points showing several wrapped-line trends in the presence of outliers. (Bottom) Log likelihood as a function of DOA. The parameter κ determines how strongly outliers are penalized.

C. Probabilistic Circular-Linear Regression

Consider the case of fitting a single wrapped line of the form (12) to an IPD dataset $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_{ft}\}$. We can score a line with slope α according to a likelihood criterion such as:

$$\mathcal{L}\left(\boldsymbol{\Delta}\,;\,\alpha\right) = \prod_{f=1}^{D}\prod_{t=1}^{T} p\left(\boldsymbol{\delta}_{ft}\,;\,\psi(\alpha f),\kappa\right) \quad , \tag{15}$$

where the probability distribution $p\left(\boldsymbol{\delta}_{ft}\,;\,-,-\right)$ is arbitrarily chosen to be von Mises for the sake of discussion.

An example of multimodal circular-linear data and the corresponding likelihood function (15) for three values of κ are shown in Fig. 3. Roughly half of the data can be considered outliers. Nevertheless, it is clear that peaks of (15) correspond to the orientations of the wrapped lines.

In Section VI, we will see how to separate the speakers given estimates of their IPD lines. What remains is a robust and efficient algorithm for estimating the slopes $\boldsymbol{\alpha}$.

IV. Fitting a Mixture of Wrapped Lines with EM

We have reduced the source separation problem to one of multimodal circular-linear regression. We now view this as a parameter estimation problem with latent variables and apply the Expectation-Maximization (EM) framework [29]. The observed variables are the IPD features $\boldsymbol{\Delta}$, the hidden variables are the TF bin labels $\mathbf{z}$, and the unknown parameters are the slopes $\boldsymbol{\alpha}$, variances $\boldsymbol{\sigma}^2$, and weights $\boldsymbol{\pi}$. In this section, we review the EM algorithm for fitting a mixture of wrapped Gaussians and extend it to cluster across frequencies.

Algorithm 1 EM to fit Mixture of Wrapped Gaussians

E step

$$\eta_{tjl} = \frac{\mathcal{N}\left(\delta_t\,;\,\widehat{\mu}_j+2\pi l\,,\,\widehat{\sigma}_j^2\right)\,\widehat{\pi}_j}{\sum_{j=1}^{K}\sum_{l=-\infty}^{\infty}\mathcal{N}\left(\delta_t\,;\,\widehat{\mu}_j+2\pi l\,,\,\widehat{\sigma}_j^2\right)\,\widehat{\pi}_j}$$

M step

$$\widehat{\mu}_j = \frac{\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}(\delta_t-2\pi l)\,\eta_{tjl}}{\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\eta_{tjl}}$$

$$\widehat{\sigma}_j^2 = \frac{\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}(\delta_t-\widehat{\mu}_j-2\pi l)^2\,\eta_{tjl}}{\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\eta_{tjl}}$$

$$\widehat{\pi}_j = \tfrac{1}{N}\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\eta_{tjl}$$

A. Clustering in a single frequency band

Consider a dataset of T samples $\delta_t \in \mathbb{S}^1, t = 1,\ldots,T$ drawn i.i.d. from a mixture of wrapped Gaussians (MoWG) [32]. The MoWG pdf, defined over $\mathbb{S}^1$, is given as:

$$p\left(\delta\,;\,\boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\pi}\right) = \sum_{j=1}^{K}\pi_j\sum_{l=-\infty}^{\infty}\mathcal{N}\left(\delta\,;\,\mu_j+2\pi l,\sigma_j^2\right) \ . \tag{16}$$

A standard derivation leads to the EM procedure summarized in Algorithm 1. The posterior:

$$\eta_{tjl} = p\left(z_{jl}\,|\,\delta_t\,;\,\mu_j,\sigma_j^2,\pi_j\right) \ , \tag{17}$$

represents the assignment probability of the t^{th} sample to the l^{th} component of the j^{th} WG. We truncate the infinite summations to 3 terms centered at $l = 0$ for tractability.

Clustering in each frequency band individually fails to capture the wrapped-linear structure of the dataset, making it unclear how to group the clusters across frequencies according to speaker identity [33]. Thus, we modify this algorithm to perform the clustering jointly.

B. Clustering across frequency bands

To incorporate the linear trend of the IPD data across frequency, we reparameterize the model from the previous section in terms of the slopes α_j:

$$\mu_{jf} = \alpha_j f \ . \tag{18}$$

In each frequency band, we have a mixture of WG distributions and the means for each source are tied together across frequency. For tractability, we assume statistical independence between the univariate WGs. The joint pdf of these mean-locked mixtures of wrapped Gaussians (ML-MoWG) is:

$$p\left(\boldsymbol{\delta}\,;\,\boldsymbol{\alpha},\boldsymbol{\sigma}^2,\boldsymbol{\pi}\right) = \prod_{f=1}^{D}\sum_{j=1}^{K}\pi_{jf}\sum_{l=-\infty}^{\infty}\mathcal{N}\left(\delta_f\,;\,\alpha_j f+2\pi l\,,\,\sigma_{jf}^2\right). \tag{19}$$

Algorithm 2 EM to fit Mean-Locked Mixtures of Wrapped Gaussians

E step

$$\eta_{ftjl} = \frac{\mathcal{N}\left(\delta_{ft}\,;\widehat{\alpha}_j f{+}2\pi l\,,\widehat{\sigma}^2_{jf}\right)\widehat{\pi}_{jf}}{\sum_{j=1}^{K}\sum_{l=-\infty}^{\infty}\mathcal{N}\left(\delta_{ft}\,;\widehat{\alpha}_j f{+}2\pi l\,,\widehat{\sigma}^2_{jf}\right)\widehat{\pi}_{jf}}$$

M step

$$\widehat{\alpha}_j = \frac{\sum_{f=1}^{D}\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\frac{f\left(\delta_{ft}-2\pi l\right)}{\widehat{\sigma}^2_{jf}}\,\eta_{ftjl}}{\sum_{f=1}^{D}\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\frac{f^2}{\widehat{\sigma}^2_{jf}}\,\eta_{ftjl}}$$

$$\widehat{\sigma}^2_{jf} = \frac{\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}(\delta_{ft}-\widehat{\alpha}_j f-2\pi l)^2\,\eta_{ftjl}}{\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\eta_{ftjl}}$$

$$\widehat{\pi}_{jf} = \frac{1}{T}\sum_{t=1}^{T}\sum_{l=-\infty}^{\infty}\eta_{ftjl}$$

Given a dataset of T vectors $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_t\}$, $\boldsymbol{\delta}_t \in \mathbb{S}^D$, sampled i.i.d., we can solve for the parameters using an EM algorithm [23] (see Algorithm 2). The posterior:

$$\eta_{ftjl} = p\left(z_{jl}\,|\,\delta_{ft}\,;\,\alpha_j\,,\,\sigma^2_{jf}\,,\,\pi_{jf}\right)\quad, \tag{20}$$

represents the assignment probability of the $(f,t)^{\text{th}}$ data point to the l^{th} component of the j^{th} WG in the f^{th} frequency.

We must consider more terms in the summation over l than in Algorithm 1 since a wrapped line may cover multiple cycles (see (18)). The number of terms is chosen based on the amount of wrapping expected in the highest frequency band for the largest time delay. Algorithm 2 is easily extended for multiple microphone pairs by using multivariate WGs in (19).

C. Discussion of the EM approach

The wrapped and noisy nature of IPD data leads to the presence of many local optima. Since EM performs a local optimization, it will most likely converge to a solution that doesn't correspond to the true source DOAs. Fig. 4 shows a ML-MoWG fit to IPD data. The features were extracted from a synthetic mixture of two speakers in a 2D room of size 5×5 meters.[2] The data is colored according to the posteriors from the last iteration of EM.[3] Fig. 5 depicts contours of the likelihood as a function of the slopes. In this visualization, the variances and weights were held fixed at 0.1 and 0.5, respectively. Each trace shows the progress of one run of EM from a random initialization.

EM has the advantage of explicitly modeling the wrapped nature of the data. The clustering will generally succeed if the initial parameters are close enough to the correct solution. In the next section, we present a fast method to find the slopes that serves as an initialization for EM.

V. Multimodal regression by random sampling

We now describe a fast method for clustering noisy IPD data based on the RANSAC algorithm [28], [34].

[2]The T_{60} reverberation time was 130 milliseconds.

[3]EM converged when the log likelihood improved by less than 10^{-4} % between iterations.

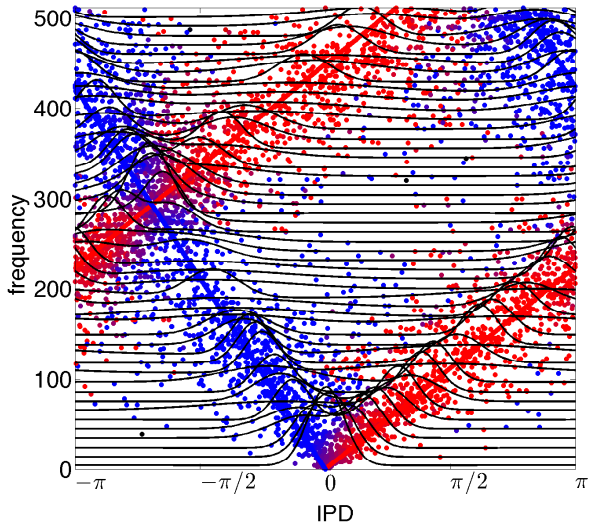


Fig. 4. Two-component, mean-locked mixtures of wrapped Gaussians fit to IPD data with EM. The data is colored according to its posterior probability and 50 of the mixtures are superimposed.

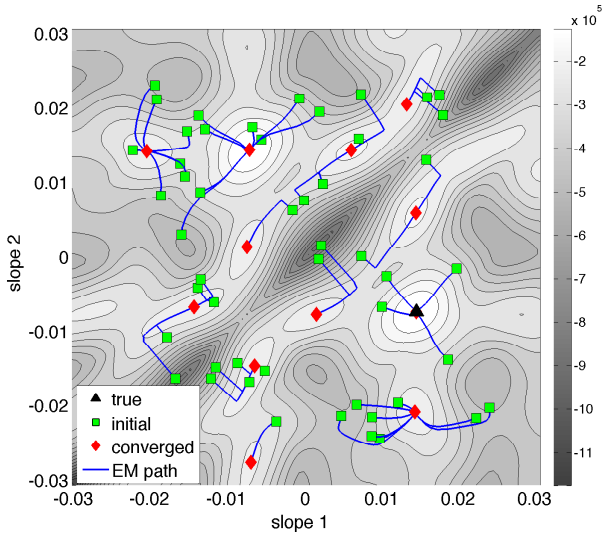


Fig. 5. Log likelihood contours for the IPD data in Fig. 4 over the slopes with the variance and mixing weights held fixed. Each trace shows the path of EM to a local maximum. A good initialization is necessary.

A. RANSAC

RANSAC is a hugely important method in the computer vision literature for estimating a simple model from a dataset with a substantial proportion of outliers. If the model can be fully described by a small set of points, one simply needs to find such a set in the data to recover the parameters. In this paper, we are interested in fitting a line that passes through the origin in IPD space. Thus, we only need one point to fully specify the model. Candidates (samples) are chosen at random from the dataset $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_{ft}\}$ such that at least one is an inlier of the true model with high probability. The inlier criterion is chosen to reflect similarity between data points.

We must sample a sufficiently high number M of candidates to ensure a good fit. This is given by the expected number of trials $\mathrm{E}[t]$ until an inlier is chosen. If the proportion of

Algorithm 3 Sequential RANSAC for Wrapped Line Fitting

Inputs: $\boldsymbol{\Delta} = \{\boldsymbol{\delta}_i\} : N$ feature vectors
K : number of wrapped lines to fit
Output: $\widehat{\boldsymbol{\alpha}} = \{\widehat{\alpha}_j\} : K$ slopes
$\mathbf{Y} = M$ samples from $\boldsymbol{\Delta}$ selected uniformly at random
$\mathbf{I} = \mathbf{0}^{N \times M}$
for $m = 1 : M$ **do**
Fit line with slope α_m to Y_m
$\mathbf{I}(i, m) = 1$ if $\boldsymbol{\delta}_i$ is inlier of m^{th} line (see (21))
end for
$\widehat{\boldsymbol{\alpha}} = \{\ \}$
$A = \{1, \ldots, N\}$
for $j = 1 : K$ **do**
$\widehat{m} = \underset{m}{\operatorname{argmax}} \sum_{i \in A} \mathbf{I}(i, m)$
$\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}} \cup \alpha_{\widehat{m}}$
$A = A \setminus \{i : \mathbf{I}(i, \widehat{m}) = 1\}$
end for

inliers in the dataset is p and we need one data point to fit a model, it can be shown [28] that $\mathrm{E}[t] = p^{-1}$. In practice, M is overestimated for robustness (e.g. $M = 10\, p^{-1}$).

B. Sequential RANSAC

Efficient, sequential variants of RANSAC have been proposed to identify multiple planar homographies for a stereo imaging application [35], [36]. As discussed in [7], we can apply a similar approach to cluster the IPD data. We adopt the same probabilistic model as in (19), replacing the WG with the von Mises (vM) for the sake of convenience. The procedure is summarized in Algorithm 3, where the data is indexed by i rather than an $\{f, t\}$ pair for clarity. The matrix I indicates whether each sample is an inlier of each candidate line and A indicates the set of samples that have not yet been counted as inliers of any line. In practice, M is scaled up by the number of sources K.

C. Example

An illustration of Algorithm 3 as applied to IPD data is shown in Fig. 6. Fig. 6(a) shows five RANSAC samples chosen uniformly at random from the data. Fig. 6(b) shows the line candidates corresponding to these samples and their inlier counts. The orange line is chosen and removed along with its inliers. This process is repeated to find the next best candidate, the yellow line, as shown in Fig. 6(c).

D. Why sequential RANSAC works

We have found that sequential RANSAC works very well for a wide range of conditions. In [7], we gave the example of a stereo recording in a stairwell whose T_{60} reverberation time was 1.5 seconds. Even though outliers made up roughly 65% of the data, the line-fitting procedure was still successful. We can understand this by considering the original probabilistic model presented in Section III-C.

When RANSAC samples are drawn from the dataset $\boldsymbol{\Delta}$, they are effectively sampled from the likelihood function shown in Fig. 3. We expect to draw candidates most frequently from high-density regions in the likelihood function. This is why so few samples are required to fit the IPD lines.

Although this method is fast and robust to outliers and wrapping issues, it guarantees only that the solution will be *near* a global optimum with high probability. It is therefore beneficial to use RANSAC as an initialization scheme for the EM algorithm. From this starting point, EM will refine the parameters of the wrapped lines and converge to the nearest maximum of the log likelihood.

We found that this approach works well on synthetic data sampled from a ML-MoWG with a portion of the samples replaced by uniform noise (to simulate outliers). However, this is not generally the case when working with IPD features extracted from a noisy, reverberant audio mixture due to model mismatch. This can be remedied by a number of adjustments that are discussed in the experiments section.

VI. Blind separation of stationary sources

We review the BSS method in [7] based on sequential RANSAC for stereo unmixing and elaborate on how this is extended when 3 or more microphones are available. Phase difference features $\boldsymbol{\delta}_{ft}$ are constructed as in (13). Sequential RANSAC is then applied to fit K wrapped lines. The $(f, t)^{\text{th}}$ point is considered an inlier of the j^{th} line if δ_{ft} is within $\pm\tau$ of the mean $\mu_{jf} = \psi(\alpha_j f)$. Since the IPDs live in a circular domain, this is equivalent to the criterion:

$$\cos(\delta_{ft} - \mu_{jf}) \geq \cos(\tau) \quad . \tag{21}$$

Note that $\cos(-)$ is a measure of proximity (as opposed to distance). The optimal choice of threshold depends on the recording environment (i.e. how noisy the IPD features are). We found that $\tau = \pi/8$ was an effective default value (a more thorough discussion can be found in [37]).

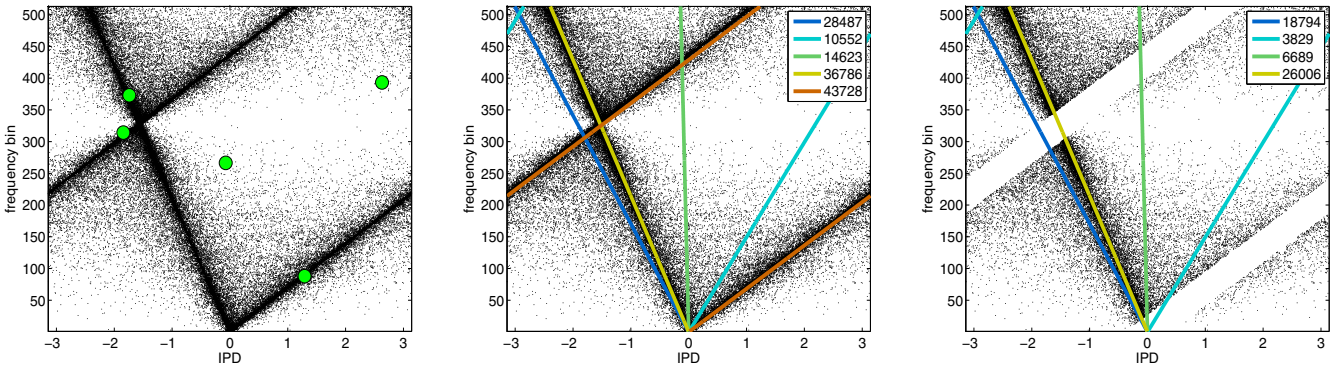
To recover the K source signals, we apply TF masks to the mixture STFT and transform the result to the time domain with the overlap-add algorithm. The mask weights in each bin are calculated via the posterior probabilities:

$$w_{ftj} = \frac{p(\boldsymbol{\delta}_{ft};\, \mu_{jf}, \kappa)}{\sum_{j=1}^{K} p(\boldsymbol{\delta}_{ft};\, \mu_{jf}, \kappa)} = \frac{e^{\kappa \cos(\delta_{ft} - \alpha_j f)}}{\sum_{j=1}^{K} e^{\kappa \cos(\delta_{ft} - \alpha_j f)}} \quad . \tag{22}$$

These probabilities represent the soft assignment of the $(f, t)^{\text{th}}$ bin to the j^{th} source. More aggressive separation is achieved (at the cost of artifacts) with a larger κ. In the limit as $\kappa \to \infty$, (22) reduces to a binary mask where each bin contributes to the reconstruction of only one source:

$$\forall\, f, t \quad w^b_{ftj} = \begin{cases} 1 & \text{if} \quad w_{ftj} = \max_l\ w_{ftl} \\ 0 & \text{else} \end{cases} \quad . \tag{23}$$

We can also apply this technique with more than two channels by extending the IPD feature vectors as in (14). The higher-dimensional data has multiple circular axes instead of



(a) (b) (c)

Fig. 6. (a) IPD data with 5 RANSAC samples overlaid. (b) First iteration of sequential RANSAC showing candidate wrapped lines and their inlier counts. (c) Second iteration of sequential RANSAC after removal of the inliers of the first model.

one. This can never decrease the inter-cluster distances and may increase them substantially, resulting in better clustering and separation. We calculate inliers by generalizing (21):

$$\sum_{i=1}^{C-1} \cos\left(\delta_{ft}(1, i+1) - \mu_{jfi}\right) \geq (C-1)\cos(\tau) \quad , \qquad (24)$$

where $\mu_{jfi} = \psi(\alpha_{ij} f)$ is the value of the j^{th} wrapped line in the f^{th} frequency band and i^{th} circular axis. α_{ij} denotes the j^{th} slope in the i^{th} circular axis. This criterion assumes that the error is measured with a multivariate von Mises distribution [38] with independent components. The mask weights are also generalized via the multivariate von Mises:

$$w_{ftj} = \frac{\prod_{i=1}^{C-1} e^{\kappa \cos(\delta_{f,t}(1,i+1) - \alpha_{ij} f)}}{\sum_{j=1}^{K} \prod_{i=1}^{C-1} e^{\kappa \cos(\delta_{f,t}(1,i+1) - \alpha_{ij} f)}} \quad . \qquad (25)$$

VII. Concurrent Source Tracking and Separation

We have shown how sequential RANSAC can be used to estimate multiple wrapped line slopes from IPD features. This assumed that the sources were physically stationary relative to the microphone array. If the sources are moving, the slopes will change over time. We assume that the source positions don't change too quickly between STFT frames. The Bayesian filtering framework is quite popular in this context as it provides a method for recursively estimating an unobserved quantity over time from noisy measurements [39]. We will track the sources' DOAs instead of their slopes as this leads to a reduction in the variance of the filter.[4] RANSAC will be used on a short-term basis to generate DOA votes that act as measurements in the tracking algorithm. We implement the tracking with a Factorial Wrapped Kalman Filter (FWKF). Details of the WKF and FWKF can be found in [23].

A. Conversion from IPD line slope to DOA

We will relate the slope α of an IPD line to the azimuthal direction-of-arrival (DOA) θ of a sound source. We know from (12) that the IPD line slope α is linearly related to the inter-channel delay, $e_{12} = d_2 - d_1$, by:

$$\alpha = -\frac{\pi}{D} e_{12} \quad . \qquad (26)$$

Slopes are converted to time delays and the least-squares method [40] can be applied to estimate the DOA. Note that this requires knowledge of the array geometry.

B. Factorial Wrapped Kalman Filter

We can track a speaker in azimuthal DOA space in the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ with a 2-microphone array and the standard Kalman filter. When tracking on the unit circle with 3 microphones, an issue arises since the DOAs $-\pi$ and π are identical. The Wrapped Kalman Filter (WKF) [22] deals with this in a statistically-grounded way by representing the DOA distribution as a wrapped Gaussian.

When multiple sources are present, we can apply the Factorial WKF (FWKF) [23]. The FWKF consists of multiple WKFs running in parallel, where observations are probabilistically associated with each WKF according to the likelihoods under each emission model.[5]

C. Blind separation of moving sources

We track moving speakers with a FWKF using DOA votes extracted with RANSAC. These votes act as measurements to "steer" the filters. At time t, we compute an IPD feature δ_{ft} for each frequency bin and determine if each feature is an inlier of each of the wrapped lines using (21). The means μ_{jf} are determined by calculating the inter-channel delays implied by the source DOAs and relating the delays to the means with (26) and (18). A feature that is not an inlier of at least one IPD line is removed. The remaining features are transformed to DOA votes using the method described in Section VII-A and passed on to the FWKF. The FWKF computes a final measurement for each source via weighted averages of the DOA votes (see [23] for an extensive description of the FWKF).

[4]The DOA and slopes reside in $\mathbb{S}^1$ and $\mathbb{R}^{C-1}$, respectively [22].

[5]Methods for assigning observations to tracks can be found in the literature on Probabilistic Data Association [41] and Multiple Hypothesis Tracking [42].

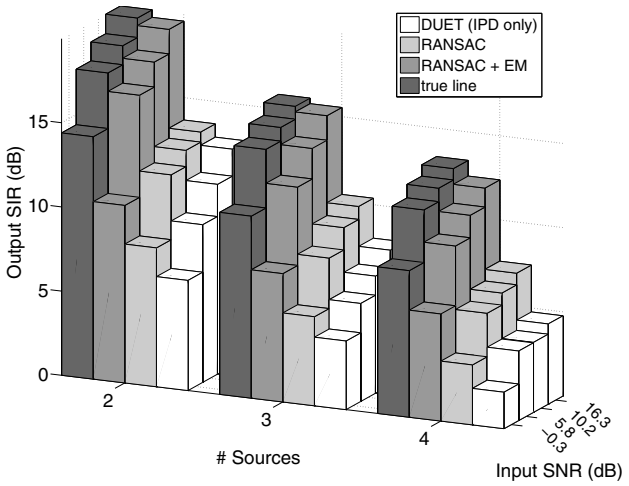


Fig. 7. Average source separation performance of IPD-based methods for stationary speakers in reverberant babble noise averaged over 100 trials. The input SNR is the ratio of target signals energy to babble noise energy.

The track estimates from each WKF dictate what IPD lines to use for calculating the mask weights in the t^{th} frame of the mixture STFT. Indeed, any method for adapting the IPD lines as the sources move is applicable here. For example, we could track the slopes of the IPD lines directly. However, this is undesirable for two reasons: (1) we would be tracking in a higher-dimensional space, incurring greater variance in the DOA estimator and (2) we are not guaranteed to maintain slope estimates that are physically consistent with a DOA. The second issue is important if we want to use the DOA tracks for more than just source separation. Thus, we chose the FWKF for its statistical grounding, interpretability, and effectiveness.

VIII. Experiments

We demonstrate the utility of the proposed algorithms through a number of experiments.[6] We first apply RANSAC and RANSAC+EM to separate stationary speakers from multichannel mixtures. Then, we introduce a FWKF to simultaneously track and separate moving speakers. In all experiments, the number of speakers K is assumed to be known and the audio is resampled to 16 kHz.

A. Separation of stationary sources

We ran source separation experiments in a 2D, simulated room with walls of length 5 meters using 2- to 3-second utterances from the TSP speaker database [43]. The array consisted of three microphones placed in a right triangle formation with sides of 8 centimeters and was positioned in the middle of the room. The speakers were placed at random on a unit circle centered at the array. The speaker DOAs were chosen from non-overlapping sections of equal length on the circle. We applied the image method [44] to simulate room reverberation with a T_{60} time of 0.55 seconds. The speakers signals were roughly equal in energy to give a 0-dB mixture. Twenty other speakers were placed uniformly at random in the room to simulate babble noise with the condition that they lie at least 1 meter away from the array. STFTs were calculated with a window size of 1,024 samples and 1/4 hop size. At a sampling rate of 16 kHz, each frame consists of 64 milliseconds of audio.

[6]Demo code is available at http://cal.cs.illinois.edu/~johannes/research/RANSAC_MoWG_bss.zip

TABLE I: BSS Eval [45] metrics for "dev3" dataset from SiSEC 2013 for two T_{60} reverb times: 130 ms, 380 ms. Mixtures are 10-second, 3-channel recordings of 4 speakers.

Method	SDR (dB)	SIR (dB)	SAR (dB)	Time (s)
RANSAC	0.70, -1.95	1.35, -1.19	**12.74**, **10.90**	**1.6**
RANSAC+EM	**2.23**, -2.36	**7.26**, **1.46**	6.86, 3.92	9.79
[31]	-1.38, -2.83	-0.88, -2.09	12.60, 10.37	15.9
[46]	0.9, **-1.2**	3.9, -0.8	9.0, 8.5	14,400

We found that naively applying Algorithm 2 with the slopes initialized by RANSAC led to unsatisfactory results. This is due to mismatch between the statistics of the IPD data and the probabilistic model given in (19). However, RANSAC reliably provides accurate slope estimates. In addition, we found that we get the best separation results by holding those slopes, and therefore the means in (18), fixed. Since the means are fixed, few iterations are required to adapt the variances and weights (no more than 10). This approach is highly efficient since RANSAC runs several times faster than real-time and the EM iterations can be parallelized across frequencies. We included a uniform component with half the weight of the mixture to assist in rejecting outliers. This further improved the separation quality by 1-4 dB.

Fig. 7 shows the separation results with soft masking for a DUET-like approach, RANSAC, EM initialized with RANSAC, and the true wrapped line model evaluated with the BSS Eval toolbox [45]. The mask weights were derived via (25) with $\kappa = 5$ for the first two and via the posterior probabilities from the E step of Algorithm 2 for the latter two. The DUET approach estimates the slopes by finding the K largest peaks of a smoothed histogram of frequency-normalized IPD features. We improved the robustness of the methods by weighting each IPD feature by the magnitude of its associated TF bin. In the DUET and RANSAC methods, this is done by replacing each sample's contribution to the histogram/inlier count by its corresponding weight. In the EM algorithm, the weights are multiplied with the posteriors in the E step. In the DUET method, spurious peaks are detected as a result of spatial aliasing [5]. In contrast, the proposed methods explicitly model the wrapped nature of the data, leading to better separation performance.

In addition to these synthetic experiments, we compared the proposed algorithms with two others using the "dev3" dataset from the 2013 Signal Separation and Evaluation Campaign (SiSEC).[7] It consists of 4 10-second-long reverberant mixtures of 4 speakers with T_{60} times of 130 or 380 milliseconds. Results[8] for the source spatial image estimation task are summarized in Table I. In general, either of the proposed

[7]For more information, see https://sisec.wiki.irisa.fr/tiki-index.php.

[8]See [45] for definitions of the Signal-to-Distortion (SDR), Signal-to-Interference (SIR), and Signal-to-Artifact (SAR) Ratios.

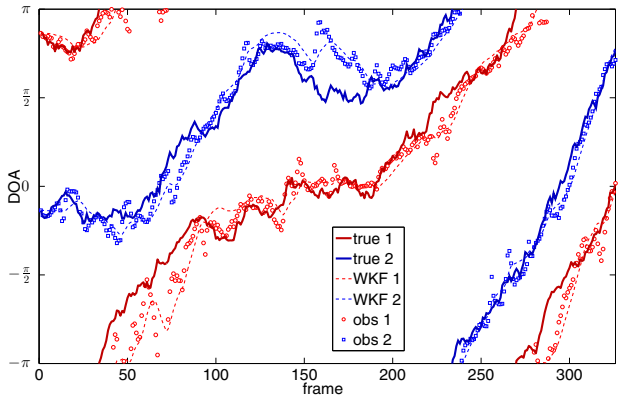


Fig. 8. DOA paths computed from a recording of two speakers using measurements selected with RANSAC. We can separate the speakers despite heavy reverberation and interfering babble noise.

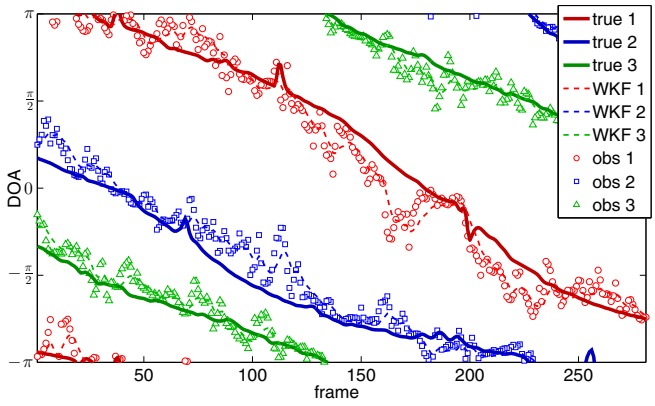


Fig. 9. DOA paths found by the FWKF for a real mixture of three speakers.

algorithms outperforms the others according to each metric. The advantage of adapting the model parameters with EM is evident in the SIR numbers, which are vastly improved over those of the RANSAC method alone.

The algorithm in [31], called MLESAC-F, identifies the IPD lines by searching over all valid inter-channel time delays. MLESAC-F appears similar to RANSAC+EM, but there are key differences. First, RANSAC+EM constructs a TF mask that is adaptive to the data. This is not done in [31]. Second, MLESAC-F identifies the source DOAs by computing perpendicular distances between IPD features and wrapped line candidates. This requires that all IPDs be replicated every $\pm 2\pi$ radians. The added computations cause MLESAC-F to run slower than real time (see Table I). In contrast, the proposed methods explicitly model the directional statistics of the IPDs, avoiding the replication step. Finally, MLESAC-F requires knowledge of the array geometry to search over all valid time delays, whereas the proposed methods do not.

B. Separation of moving sources

We ran experiments in a simulated environment identical to that used in section VIII-A to test the ability of the FWKF to adapt the wrapped line models over time. The array was placed in the (horizontal) plane of the speakers. The speaker locations evolved according to a wrapped dynamical system (WDS) [22], which models DOA and angular velocity. We included an additional component for the distance from the array. Handling cross-over in the DOA tracks is beyond the scope of this paper, so we included a repulsive step to prevent it. In these experiments, we used the masking procedure described in Section VI with binary masking.

Fig. 8 shows a trial with two speakers using three frames to extract DOA votes. The target-to-ambient-speech SNR was 6.16 dB. The T_{60} reverberation time was 0.42 seconds with a Direct-to-Reverberant Ratio [47] of 10.35 dB. Under these noisy, reverberant conditions, the tracking is challenging during pauses in the target speech and when a speaker moves far from the array. Despite this, the FWKF (equipped with RANSAC) tracks the speakers to achieve an output SIR comparable to the stationary case (18.04 dB). In contrast, the batch RANSAC method achieves an SIR of 0.79 dB. This shows the importance of adapting the IPD lines over time.

We performed a similar experiment with two and three real speech sources. The array consisted of three omnidirectional Behringer ECM8000 microphones placed 1.5 meters above the ground with the same configuration as in previous experiments. Sentences from the TSP database were played through a loudspeaker as it was swept around the array at a radius of approximately 1 meter. We recorded each source as well as ambient speech separately and added the signals together such that the target-to-ambient-speech SNR was 6 dB. The T_{60} time was approximately 0.25 seconds for the target speakers and 0.45 seconds for the ambient speakers. We estimated the ground truth DOAs with the IPDs from the individual speaker recordings as the peak of (15) in each frame.

The tracking results for the three-speaker experiment are shown in Fig. 9. Using the individual speaker recordings as a reference, we found that the SIR for the ground truth DOA parameters and the FWKF and batch RANSAC methods were 10.5, 9.6, and, 2.6 dB for the two-speaker experiment and 3.2, 3.2, and -2.5 dB for the three-speaker experiment. Qualitatively, this mirrors the results of the synthetic trial. The quantitative difference is likely due to real-world nonlinearities in the IPD features [11].

IX. Conclusion

In this paper, we addressed the problem of separating speakers from a multichannel recording by regarding IPD features as samples from a circular-linear probabilistic model. We presented a novel EM algorithm to simultaneously discover the wrapped-line structure across frequency bands and the mixture distributions within each frequency band. We showed that this approach requires a good initialization to find a good solution. Thus, we introduced an approach based on the RANSAC algorithm to estimate the line slopes. We showed that RANSAC is highly robust to outliers and provides an excellent initialization scheme for EM. We then presented experiments showing that RANSAC and EM can be combined to achieve good speech separation performance at a low computational cost. Finally, we extended the proposed method to handle moving sources by tracking the speakers' DOAs

with a Factorial Wrapped Kalman Filter (FWKF) and using RANSAC as a feature pre-processor.

X. Acknowledgement

The authors would like to thank the anonymous reviewers for their helpful comments.

References

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

[2] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, p. 23672387, 2004.

[3] P. Common and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, Academic Press, 1 edition, 2010.

[4] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *International Workshop on Independence and Artificial Neural Networks*, 1998.

[5] 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.

[6] 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.

[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 (ICASSP)*, 2013.

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

[9] 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.

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

[11] J. Traa and P. Smaragdis, "Robust interchannel phase difference modeling with wrapped regression splines," *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2014.

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

[13] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian acoustic localization," *in the proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.

[14] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 549–557, 2003.

[15] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[16] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 19, no. 1, pp. 134–139, 1983.

[17] R. Roy and T. Kailath, "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[18] G-C Hsieh and J Hung, "Phase-locked loop techniques. A survey," *IEEE Transactions on Industrial Electronics*, vol. 43, no. 6, pp. 609–615, 1996.

[19] J. Estrada, M. Servin, and J. Quiroga, "Noise robust linear dynamic system for phase unwrapping and smoothing," *Optics Express*, vol. 19, no. 6, 2011.

[20] K. Mardia and P. Jupp, *Directional Statistics*, Wiley, 1999.

[21] 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.

[22] J. Traa and P. Smaragdis, "A wrapped Kalman filter for azimuthal speaker tracking," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1257–1260, 2013.

[23] J. Traa, "Multichannel source separation and tracking with phase differences by random sample consensus," M.S. thesis, University of Illinois at Urbana-Champaign, 2013.

[24] G. Welch and G. Bishop, "An introduction to the Kalman filter," Tech. Rep., University of North Carolina at Chapel Hill, 2006.

[25] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic source tracking applying Rao-Blackwellized Monte Carlo data association," *IEEE 15th Workshop on Statistical Signal Processing*, pp. 253–256, 2009.

[26] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

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

[28] 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.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[30] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[31] L. Litwic and P. J. Jackson, "Source localization and separation using Random Sample Consensus with phase cues," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 337–340, 2011.

[32] Y. Agiomyrgiannakis and Y. Stylianou, "Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 775–786, 2009.

[33] H. Sawada, R. Mukai, S. Araki, and S. Makino, *Speech Enhancement - Chapter 13: Frequency Domain Blind Source Separation*, Springer, 2005.

[34] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, Prentice Hall, 2003.

[35] 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.

[36] 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.

[37] J. Traa and M. Kim, "Phase and level difference fusion for robust multichannel source separation," *in the proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[38] K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh, "A multivariate von Mises distribution with applications to bioinformatics," Tech. Rep., University of Leeds, 2007.

[39] V. Cevher, R. Velmurugan, and J. H. McClellan, "Acoustic multitarget tracking using direction-of-arrival batches," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2810–2825, 2007.

[40] K. M. Varma, "Time delay estimate based direction of arrival estimation for speech in reverberant environments," M.S. thesis, Virginia Polytechnic Institute and State University, 2002.

[41] T. Kirubarajan and Y. Bar-Shalom, "Probabilistic data association techniques for target tracking in clutter," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 536–557, 2004.

[42] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[43] P. Kabal, "TSP speech database," 2002, Telecommunications and Signal Processing Lab, McGill University.

[44] 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.

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

[46] K. Adiloglu and E. Vincent, "Variational bayesian inference for source separation and robust feature extraction," Tech. Rep., INRIA, 2012.

[47] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.

Johannes Traa (M '13) received the B.S. degree in electrical engineering from Northwestern University, Evanston, IL, in 2011 and the M.S. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, in 2013. He is currently pursuing the Ph.D. degree in electrical and computer engineering at UIUC.

He has been a Research Assistant with Paris Smaragdis at UIUC since 2011. In the summers of 2012-2014, he interned with the theory group at Lyric Labs, Analog Devices in Boston, MA. His research interests include audio source separation and localization, sound mixture analysis with additive modeling techniques like non-negative matrix factorization, and applications of various areas of statistics (e.g. compositional, directional) to audio problems.

Paris Smaragdis (paris@illinois.edu) is an assistant professor in the Computer Science Department and the Electrical and Computer Science Department at the University of Illinois at Urbana-Champaign, as well as a senior research scientist at Adobe Research. Prior to that he was a research scientist at Mitsubishi Electric Research Labs, during which time he was selected by MIT Technology Review as one of the top 35 young innovators of 2006. His research interests lie in the intersection of machine learning and signal processing. He is a Senior Member of the IEEE.