TWO MICROPHONE BASED DIRECTION OF ARRIVAL ESTIMATION FOR MULTIPLE SPEECH SOURCES USING SPECTRAL PROPERTIES OF SPEECH

Wenyi Zhang and Bhaskar D. Rao

Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407, USA
Email: w3zhang@ucsd.edu, brao@ece.ucsd.edu

ABSTRACT

A two microphone direction of arrival (DOA) estimation technique for multiple speech sources is developed which exploits speech specific properties, namely sparsity in time-frequency (spectrum) domain. For robustness, we exploit the sparsity in the frequency domain by focusing on the spectral content concentrated in sinusoidal tracks obtained through sinusoidal modeling. When multiple speeches are mixed in the two microphone system, the inter-channel phase differences (IPD) between the dual channels on those sinusoidal tracks will be dominated by the spatial information of the most powerful source at that specific time-frequency point because of the spectrum sparsity and masking effects. Thereby, the source localization problem is turned into a clustering problem on the IPD versus frequency plot, and the generalized mixture decomposition algorithm (GMDA) is used to cluster the groups of points corresponding to multiple sources. The DOA of each source is derived from the parameters of each cluster. Experimental results conducted show the scheme to be very effective.

Index Terms— Two microphone system, direction of arrival estimation, speech, sparsity, sinusoidal modeling, generalized mixture decomposition algorithm

1. INTRODUCTION

In this paper, direction of arrival (DOA) estimation using only two microphones is considered, and methods in this context can be broadly categorized into two classes: time domain approaches and frequency domain approaches. The time domain algorithms include time domain cross correlation method, average-magnitude-difference function method, LMS-type adaptive TDE algorithm, and adaptive eigenvalue decomposition algorithm associated with blind channel identification (see [1] and refs. therein). The frequency domain algorithms include linear regression method, blind channel identification based method and the well known generalized cross-correlation (GCC) family of methods, which includes many variations such as the smoothed coherence transform, the phase transform, the maximum likelihood approach among others [2, 3, 1]. However, most of the these algorithms are based on a single source signal model and can not effectively locate multiple sources. It is shown that speech specific attributes, namely sparsity in time-frequency domain, can be utilized to locate multiple speech sources using two microphones [4, 5, 6]. For instance, in [4] the presence of gaps in the spectrum of each source at different times and frequencies is exploited and an image processing method is employed to detect vertical segments in the frequency vs. path difference plot to locate two sound sources. In [5], harmonic sound stream segregation using localization is considered and a rough localization method based on harmonic streams and inter-channel phase differences (IPD) and inter-channel intensity difference (IID) is proposed. The sparse speech assumption is explicitly used in [6] for localization and a histogram type of method is proposed to locate multiple sources through the frequency vs. DOA plot.

In this paper, we follow this tradition and propose a two microphone DOA estimation technique for multiple speech sources based on the speech's sparsity attribute. Speech can be represented by sinusoidal tracks in the time-frequency (spectrum) domain according to sinusoidal modeling [7]. An advantage of utilizing the sinusoidal tracks is that they represent regions where speech energy is concentrated leading to high signal to noise ratio data points for further analysis. When multiple speech are mixed in the two microphone/dual channel system, the inter-channel phase differences (IPD) between the dual channels on those sinusoidal tracks will be dominated by the spatial information of the most powerful source at that specific time-frequency point because of the spectrum sparsity and masking effects (Sec. 2). The error between the IPD between the dual channel signals and the IPD between a source signal's contributions at the dual channels is modeled as a random variable and a statistical signal model for the IPD error is proposed. Thereby, the source localization problem is turned into a clustering problem on the IPD vs. frequency plot. Generalized mixture decomposition algorithm (GMDA) is used to cluster the groups of points representing multiple sources. The DOA of each source is derived from the parameters of each cluster. Depending on inter-microphone spacing, spatial aliasing effect is taken into consideration by proper phase unwrapping. A minimum description length (MDL) algorithm is used to determine the number of sources. Experimental results conducted demonstrate the efficacy of the proposed method.

2. TWO MICROPHONE DOA ESTIMATION FOR MULTIPLE SOURCES

2.1. Single Source Scenario

We first describe two microphone based DOA estimation using IPD for a single source to motivate the proposed method. Assuming a far field source scenario, a simple DOA estimation algorithm using two microphones can be developed based on inter-channel phase difference (IPD). Denoting the desired source signal by $s[k]$, the dual

This material is based upon work supported by the RESCUE project under National Science Foundation Award Number 0331690 and UC Micro Grants 07-034 and 08-065 sponsored by Qualcomm Inc.

channel discrete signals $x_1[k]$ and $x_2[k]$ can be expressed as,

$$x_1[k] = s[k] + n_1[k] \tag{1a}$$

$$x_2[k] = s[k-\tau] + n_2[k] \tag{1b}$$

where $s[k-\tau]$ represents a delayed version of $s[k]$, τ is the time delay of the desired source. $n_1[k]$ and $n_2[k]$ represent ambient noise and more generally also include interference signals.

The short time discrete Fourier transform (DFT) of $x_1[k]$ and $x_2[k]$ is denoted by $X_1(\omega)$ and $X_2(\omega)$ respectively.

$$X_1(\omega) = S(\omega) + N_1(\omega) \tag{2a}$$

$$X_2(\omega) = S(\omega)e^{-j\omega\tau} + N_2(\omega) \tag{2b}$$

where ω represents angular frequency, $S(\omega)$ is the DFT of $s[n]$, $N_1(\omega)$ and $N_2(\omega)$ represent DFT of $n_1[k]$ and $n_2[k]$ respectively.

The IPD $\psi_X(\omega)$ between the two channels is,

$$\psi_X(\omega) = \angle X_1(\omega) - \angle X_2(\omega) \tag{3}$$

and is constrained to be in the range $[-\pi, \pi]$ after $mod(2\pi)$ operation. $\angle X_1(\omega)$ and $\angle X_2(\omega)$ is the phase of $X_1(\omega)$ and $X_2(\omega)$ respectively. If $N_1(\omega)$ and $N_2(\omega)$ have much smaller magnitudes than the magnitude of $S(\omega)$, then

$$\psi_X(\omega) = \omega\tau + 2\pi n + v(\omega) \tag{4}$$

where $2\pi n$ represents possible phase unwrapping. $v(\omega)$ denotes IPD error, which represents a zero mean noise term. A similar problem is considered in noncoherent detection in communication, where a Gaussian distribution is proposed to approximate the probability density function (PDF) of $v(\omega)$ at high SNR. In this work we propose to approximate the PDF of the IPD error $v(\omega)$ by a zero mean Gaussian or Laplacian random variable. Laplacian distribution is considered because of its robustness to outliers. The DOA of the desired source can be derived from the following equation,

$$\tau = d\sin\theta/c \tag{5}$$

where d represents inter-microphone distance, c the sound speed, and θ the DOA of the desired source.

Using the IPDs at different frequencies ω and different frames, linear regression method can be employed to estimate the slope of curve IPD $\psi_X(\omega)$ with respect to ω in the IPD vs. frequency plot, thereby estimating the DOA of the source. To enhance robustness, it is useful to account for the SNR at each frequency and frame and it is preferable to use only those data points with high SNR. The inclusion of only reliable points in the estimation will be an important component of the method proposed in the paper.

The above DOA estimation method based on IPD performs well when there is only one source and the SNR is high. It is unable to locate multiple sources particularly when there are multiple white sources with similar power level. Expanding the signal model shown in Eq. (2) to multiple sources, it can be readily seen that if there are two sources with similar power the IPD $\psi_X(\omega)$ in Eq. (3) does not have any simple relationship to the DOA of either source.

2.2. Multiple Source Scenario

We now develop a method for the multiple source DOA estimation that exploits source specific knowledge and attempts to retain the simplicity of the IPD technique. The speech signal has a special property, namely sparsity in time-frequency domain: **Sparsity in time domain**. Natural speech generally has many short pauses and silent segments, which may occupy more than half of the total time. **Sparsity in frequency domain**. The signal power of speech is not equally distributed across the whole frequency range. For voiced speech, the signal power is concentrated on a set of equally distributed discrete frequency points, i.e. harmonics of the pitch frequency [7].

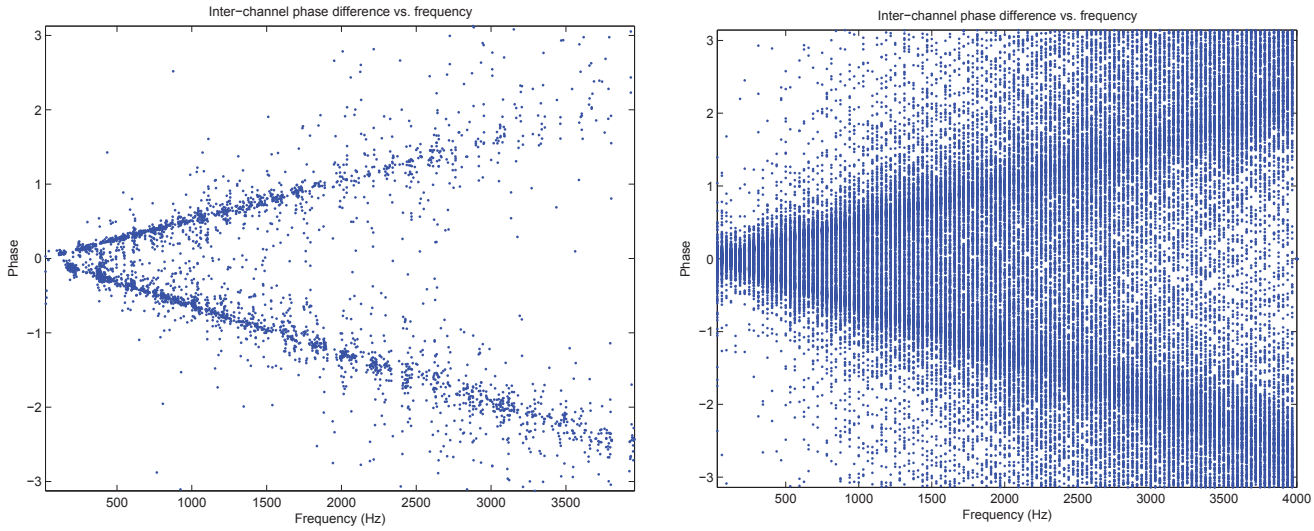
Collectively the above two attributes make the speech sound sparse in the time-frequency domain. When a recording contains multiple speeches, at a specific time-frequency point, there is a high likelihood that at most one source is dominating (in power) and the contributions from other sources is negligible. As a consequence, the IPD $\psi_X(\omega)$ (Eq. (3)) will be dominated by the IPD $\psi_S(\omega)$ of the dominating source, $\psi_S(\omega)$ is defined as the IPD between the source's contribution at the two channels, $\psi_S(\omega) = \angle S(\omega) - \angle\{S(\omega)e^{-j\omega\tau}\} = \omega\tau$. So the IPD $\psi_X(\omega)$ contains DOA information of the dominating source at that time-frequency point and can be used for DOA estimation. This is denoted as masking effect in this paper.

Speech's sparsity attribute and the masking effect combined leads to the idea of DOA estimation using only points with high local SNR in the time-frequency domain, i.e. DOA estimation based on sinusoidal modeling [7]. In [7], it is observed that speech usually has power focused on a set of discrete frequencies and can be modeled by a set of sinusoidal tracks. Sinusoidal tracks are defined to be continuous local peaks in the time-frequency domain which satisfies a set of constraints. By the sparsity and masking properties of speech signals, the sinusoidal tracks extracted from one channel of a mixed signal (can be either left or right channel) can be approximated as a disjoint union of the sinusoidal tracks from each of the different source signals, i.e. a track of the mixed signal can be associated with one of the source signals. This association is not known and will be dealt with in the next section. When two or more sources have similar power level at a time-frequency area, the interaction between source signals will cause the mixed signal $X(\omega)$ to fluctuate frequently resulting in no sinusoidal tracks in the corresponding time-frequency area. The points on the sinusoidal tracks will implicitly has high SNR.

We propose to use the IPD $\psi_X(\omega)$ between the two channels on points of the sinusoidal tracks for multiple sources DOA estimation. As an example, the IPD $\psi_X(\omega)$ on points of the sinusoidal tracks is plotted in a IPD vs. frequency plot (Fig. 1 (a)). For comparison, the IPD $\psi_X(\omega)$ on all spectrum points is also plotted (Fig. 1 (b)). The DOA of the two speech sources are 60° and -45° respectively. The inter-microphone spacing is $4cm$. There is no spatial aliasing [3] in this example. From Fig. 1 (a), it is clear there are two clusters of points which can be fitted by two lines. This represents two sources and the DOA of the two sources can be derived from the slopes of the two lines. When the IPD $\psi_X(\omega)$ for all spectrum points is plotted (Fig. 1 (b)), the cluster information is obscured and overwhelmed by noise although some cluster information can still be deduced from the plot.

In summary, the steps of the proposed dual channel multiple speech sources DOA estimation method are enumerated:

1. Calculate the time-frequency spectrum $X_1(\omega)$ and $X_2(\omega)$ of the two microphone signals using short time DFT.
2. Extract sinusoidal tracks from one of the two channels.
3. Calculate the IPD $\psi_X(\omega)$ between the two channel signals $X_1(\omega)$ and $X_2(\omega)$ on points of the sinusoidal tracks.
4. Cluster the points on the IPD vs. frequency plot, employ line fitting techniques to fit set of lines and derive DOA of sources from the slopes of the lines.



(a) IPD $\psi_X(\omega)$ on points of the sinusoidal tracks

(b) IPD $\psi_X(\omega)$ on all spectrum points

Fig. 1. Inter-channel phase difference vs. frequency, 2 sources

The last step, clustering and line fitting techniques, is discussed in the following section.

3. CLUSTERING AND LINE FITTING

3.1. Generalized mixture decomposition algorithm

This section describes a procedure that does the clustering and line fitting jointly. For this purpose, the IPD error $v(\omega)$ in Eq. (4) (the distance from a data point to the center of its underlying cluster, i.e. a line) is modeled as a Gaussian or Laplacian random variable. A mixture model is then employed to fit the data and the generalized mixture decomposition algorithm (GMDA) [8] is used to cluster the data.

Assume there are m clusters, $C_j, j = 1, .., m$, i.e. m speech sources, and m is assumed to be known. Assume there are N data points $\mathbf{y}_i, i = 1, .., N$. Each data point $\mathbf{y}_i$ is a 2-dimension vector which denotes a point on the IPD vs. frequency plot, $\mathbf{y}_i = [\omega, \psi_X(\omega)]^T$. A mixture model can be used to fit the data points. Each component of the mixture is associated with a line. The distance from each data point to its underlying line is modeled as a random variable with the PDF $p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is the parameter vector characterizing the line corresponding to the j^{th} cluster. The parameters of the mixture model are learnt from the data points using the maximum likelihood approach. The complete Expectation-Maximization (EM) algorithm herein is the generalized mixture decomposition algorithm (GMDA) [8]. The GMDA in [8] is very general and to get more specific update rules for the GMDA, an explicit form for the PDF $p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j)$ is necessary. As previously discussed in Sec. 2, an appropriate form is either a Gaussian distribution or Laplacian distribution (see [9] for more details).

In previous discussion of GMDA, the number of sources, and hence the number of clusters is assumed to be known. However, the number of clusters is not known in reality and has to be estimated from the data. This is the model selection step. Since GMDA falls into the maximum likelihood framework, minimum description length (MDL) method can be used to estimate the model order.

3.2. Clustering and line fitting under spatial aliasing scenario

In the previous discussion, it is assumed the inter-microphone spacing is small such that there is no spatial aliasing [3]. One example IPD vs. frequency plot under such scenario is shown in Fig. 1 (a). With the increasing of the inter-microphone distance spatial aliasing may exist. Fig. 2 (a) shows the IPD vs. frequency plot for the same scenario as in Fig. 1 (a) except that the inter-microphone spacing is increased to $12cm$. Recall the IPD on the original IPD vs. frequency plot is always confined to be in the range of $[-\pi, \pi]$. Two sources can still be observed in Fig. 2 (a), however, the two lines representing the two sources are broken into parallel segments because of phase wrapping effect. The two lines are easier to be observed if we do phase unwrapping and move the broken line segments parallelly and properly. This is shown in Fig. 2 (b). Note that the phase is no longer confined to $[-\pi, \pi]$.

The GMDA is now modified to handle spatial aliasing properly. Define a new IPD $\psi'_X(\omega) = \psi_X(\omega) + 2\pi n$, where n is integer and $2\pi n$ represents possible phase unwrapping. After appropriate phase unwrapping, the new data point $\mathbf{y}'_i = [\omega, \psi'_X(\omega)]^T$ will lie around its true underlying line. We propose to choose the phase unwrapping which yields the biggest probability for the observed data point. If the PDF of observing data point $\mathbf{y}_i$ given j^{th} cluster is chosen as Gaussian, denote

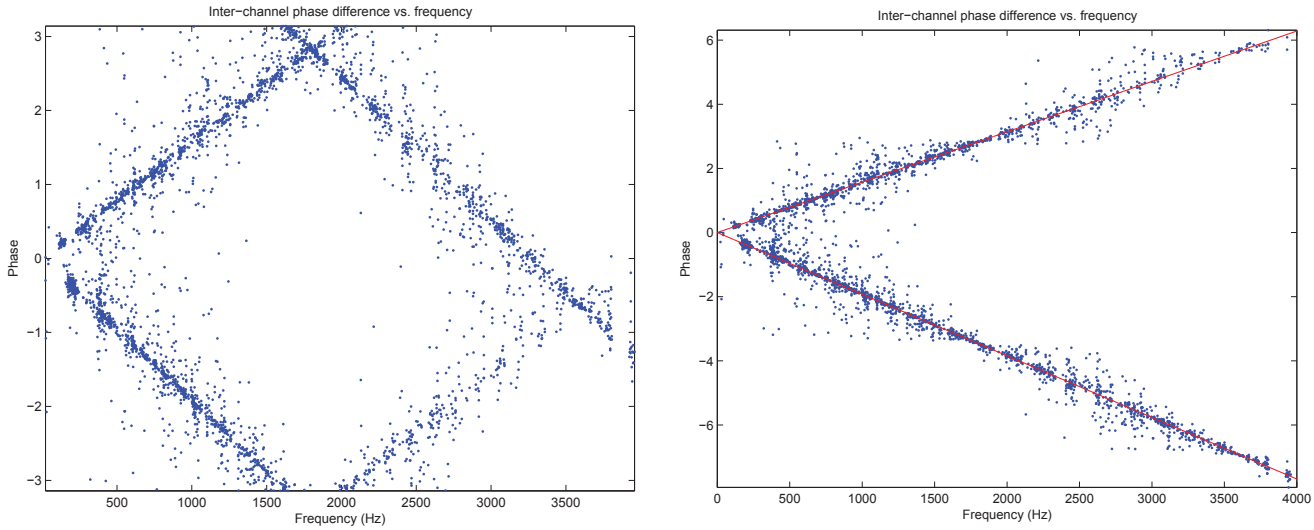
$$n^j = \arg\max_n \frac{1}{\sqrt{2\pi}\sigma_j} exp\{-\frac{(y_{i,2} + 2\pi n - \alpha_j y_{i,1})^2}{2\sigma_j^2}\}, \quad j = 1, .., m \tag{6}$$

$$J = \arg\max_j p'(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j) \tag{7}$$

then $\mathbf{y}'_i$ is chosen as $\mathbf{y}'_i = [\omega, \psi_X(\omega) + 2\pi n^J]^T$.

4. EXPERIMENTS

Simulation was conducted to demonstrate the performance of the proposed algorithm using speech sources. **Example 1**:There are 2



(a) Original (b) After adjusting phase

Fig. 2. Inter-channel phase difference vs. frequency, 2 sources, spatial aliasing scenrio

speech sources with DOA of -45° and -50° respectively and inter-microphone spacing of $4cm$. If we choose the Gaussian PDF model for the IPD error term, the estimated source number is 3 and the estimated DOAs are $(-50.00^\circ, -44.96^\circ, -54.80^\circ)$. If choose Laplacian distribution instead, the estimated source number is 2 and the estimated DOAs are $(-49.86^\circ, -45.16^\circ)$. This example demonstrates the GMDA method with Gaussian PDF model may overestimate the number of sources when the sources are spatially closely distributed. As is known, the Gaussian PDF model is sensitive to outliers in model learning. On the other hand, the Laplacian distribution for the IPD error term has a heavier tail and is more robust to outliers. In the rest of the experiments, due to space limitations, only results with the Laplacian PDF model are presented.

Example 2: There are 2 speech sources and the inter-microphone spacing is $12cm$, Table 1 illustrates the true DOA of the sources and the estimated DOAs for a variety of source configurations. The results demonstrate the proposed algorithm's to be quite reliable.

Example 3: There are three sources with DOA 60°, 0° and -45° respectively. The inter-microphone spacing is $12cm$. The estimated DOA of sources are $(60.20^\circ, 0.09^\circ, -45.09^\circ)$.

Example 4: In this experiment, the robustness of the proposed algorithm to the ambient white noise level is illustrated. High local SNR on points of the sinusoidal tracks is expected even though the global SNR across the whole spectrum might be low. Therefore, good DOA estimation would be expected even when the white noise level is high. When the SNR is higher than 10 dB, the DOA estimation is found to be quite accurate with the average DOA estimation lower than 0.5°. However, when the SNR is reduced to 0 dB, the DOA estimation error is large. This is a consequence of the failure of the sinusoidal track extraction program used to pick up the true sinusoidal tracks from the spectrum. More details and experiment results can be found in [9] .

Table 1. Estimated DOA for various configurations

Configuration index		1	2	3
True DOA	source 1	0°	40°	75°
	source 2	5°	45°	80°
Estimated DOA	source 1	0.08°	40.13°	75.15°
	source 2	4.85°	44.78°	79.74°

5. REFERENCES

[1] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. Appl. Signal Process.*, vol. 2006, no. 1, pp. 170–170, 2006.

[2] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 24, pp. 320–327, Aug 1976.

[3] H. L. V. Trees, *Optimum Array Processing*. Wiley-Interscience, 2002.

[4] D. Banks, "Localisation and separation of simultaneous voices with two microphones," *Communications, Speech and Vision, IEE Proceedings I*, vol. 140, no. 4, pp. 229–234, Aug 1993.

[5] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Commun.*, vol. 27, pp. 209–222, 1999.

[6] C. Liu, B. C. Wheeler, W. D. O'Brien, Jr., R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *Acoustical Society of America Journal*, vol. 108, pp. 1888–1905, Oct. 2000.

[7] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 34, pp. 744–754, Aug 1986.

[8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.

[9] W. Zhang and B. D. Rao, "A two microphone based approach for direction of arrival estimation of multiple speech sources," *submitted to Audio, Speech and Language Processing, IEEE Transactions on*.