DIRECTION OF ARRIVAL ESTIMATION FOR MULTIPLE SPEAKERS USING TIME-FREQUENCY ORTHOGONAL SIGNAL SEPARATION

Mikael Swartling, Nedelko Grbić, Ingvar Claesson

Department of Signal Processing, School of Engineering, Blekinge Institute of Technology
Box 520 SE-37240 Ronneby, Sweden
msw@bth.se, ngr@bth.se, icl@bth.se

ABSTRACT

This paper presents a new approach for multiple speaker DOA estimation using an array of microphones. The method relies on the fact that multiple independent speakers have a small overlap in the time-frequency domain, i.e. the individual signals are almost W-disjoint orthogonal. By introducing a time-frequency mask and by continuously track the set of time-frequency points corresponding to each individual speech signal, a single source DOA estimation algorithm is used to find the DOA for each separated signal. This approach does not limit the solution to cases where the number of sensors exceeds the number of sources. Real room recordings are used to evaluate the performance of the method where source movements are also included.

1. INTRODUCTION

This paper presents and evaluates a new method consisting of a combination of existing techniques used to determine the angle of arrival of multiple concurrent speech sources with respect to a microphone array. The method involves three steps:

1. using a blind signal separation algorithm to separate the different speech sources into sets of mixtures, each containing a single source,
2. using conventional single source methods for estimating time difference for each set of mixtures, and
3. filter angle estimates using a one-step prediction Kalman filter.

By preprocessing the mixtures with a blind signal separation algorithm, the problem of delay estimation is reduced from finding multiple delays in one set of mixtures, to finding single delays in several sets of signals.

The goal of blind signal separation is to separate a set of *unknown* signals, or sources, from a set of *known* mixtures. The mixtures are typically the output from a sensor array, where the different sensors receives different mixtures of the source signals. The term “blind” in this context means [1]

1. the source signals are not observed, and
2. no information is available about the mixing system.

To compensate the lack of information about the sources, their propagation to the sensor array and the mixing system, some assumptions must be made about the sources being separated. Such assumptions can be that sources must be statistically independent, or, as in this paper, that sources must be W-disjoint orthogonal.

One recently developed algorithm for blind signal separation is DUET, *Degenerate Unmixing and Estimation Technique* [2]. This algorithm can separate more sources than mixtures, refered to as *degenerate demixing*. Degenerate demixing is challenging in that the mixing matrix is not invertible, and traditional algorithms based on estimating the inverse of the mixing matrix does not work.

To estimate time delay a generalization of the *Generalized cross correlation* method [3] is used. The generalization extends the method to include more than two sensor signals. The Generalized cross correlation is a correlation based method, which involves maximizing the cross correlation of all sensor signals. Given two signals, where one is a time shifted version of the other, maximum cross correlation occurs at the point which corresponds to the time shift.

The delay estimates, or the corresponding angle of arrival estimates, are filtered to reduce variance. As this paper focus on speech sources, it is assumed that the source locations are constant within a small enough time frame to allow the filter to reduce noise variance, but still keep up with changes due to actual movement of the source.

2. BLIND SIGNAL SEPARATION

The recently developed algorithm for blind signal separation, DUET [2], is used in this paper. A modification making it suitable for online real-time applications is presented in [4]. The algorithm relies on the assumption that the sources are W-disjoint orthogonal.

2.1. W-disjoint orthogonality

Two signals $x_1(t)$ and $x_2(t)$ are W-disjoint orthogonal, if, for a given window function, the support of the windowed Fourier transform of $x_1(t)$ and $x_2(t)$ are disjoint sets. The windowed Fourier transform of $x_n(t)$ is defined as

$$\mathscr{F}^W\left[x_n(\cdot)\right](\omega,\tau) = \int_{-\infty}^{\infty} W(t-\tau)x_n(t)e^{-j\omega t}dt, \tag{1}$$

denoted in this paper as $S_n(\omega,\tau)$. The W-disjoint orhogonality can then be stated as

$$S_1(\omega,\tau)S_2(\omega,\tau) = 0, \qquad \forall\omega,\tau. \tag{2}$$

In practice, however, equation (2) is rarely satisfied exactly. Instead, the term *approximately W-disjoint orthogonal* is introduced, which represents the level of orthogonality of sources. In [4], it is shown that independent speech signals can be considered to be almost W-disjoint orthogonal.

2.2. Mixing parameter estimation

The original algorithm assumes a signal model where the relative difference between a source signal received by two sensors is only a

1-4244-0469-X/06/$20.00 ©2006 IEEE

scale factor and a time delay, expressed as

$$\begin{aligned}x_1(t) &= \sum_{n=1}^{N} s_n(t)\\ x_2(t) &= \sum_{n=1}^{N} a_n s_n(t-\delta_n)\end{aligned} \tag{3}$$

where N is the number of sources, a_n and δ_n is the relative attenuation and time delay, respectively, between the two sensors for source n at the sensor pair. In matrix form, this can be expressed as

$$\begin{bmatrix}X_1(\omega,\tau)\\ X_2(\omega,\tau)\end{bmatrix} = \begin{bmatrix}1 & \cdots & 1\\ a_1e^{-j\omega\delta_1} & \cdots & a_Ne^{-j\omega\delta_N}\end{bmatrix}\begin{bmatrix}S_1(\omega,\tau)\\ \vdots\\ S_N(\omega,\tau)\end{bmatrix}. \tag{4}$$

Under the assumption that the sources are W-disjoint orthogonal, that is, that at most one source is active at any time-frequency point (ω,τ), the equation can be rewritten as

$$\begin{bmatrix}X_1(\omega,\tau)\\ X_2(\omega,\tau)\end{bmatrix} = \begin{bmatrix}1\\ a_ne^{-j\omega\delta_n}\end{bmatrix}\begin{bmatrix}S_n(\omega,\tau)\end{bmatrix}, \qquad n\in[1\ldots N] \tag{5}$$

where n indicates the single active source at the corresponding time-frequency point (ω,τ).

In [4], a maximum likeliehood cost function is derived. This cost function is minimized with a gradient based search method in order to find the mixing parameters. Mixing parameters are updated as

$$\begin{aligned}a_n(k) &= a_n(k-1) - \mu\frac{\partial J(\tau)}{\partial a_n}\\ \delta_n(k) &= \delta_n(k-1) - \mu\frac{\partial J(\tau)}{\partial \delta_n}\end{aligned} \tag{6}$$

where μ is the learning rate, $J(\cdot)$ is the cost function and k is a time index.

Some modifications for small arrays are made to the original algorithm in order to improve the performance. Assuming that the relative attenuation mixing parameter is unity, there is no need to track the attenuation mixing parameter, and the expression for the partial derivative for updating the delay mixing parameter is simplified. In this case, the attenuation mixing parameter is ignored, and the delay mixing parameter is updated as in (6), where

$$\frac{\partial J(\tau)}{\partial \delta_n} = \sum_{\omega}\frac{-\omega e^{-\lambda\rho_n}}{\sum_{m=1}^{N}e^{-\lambda\rho_m}}\Im\left[X_1(\omega,\tau)X_2^*(\omega,\tau)e^{-j\omega\delta_n}\right] \tag{7}$$

and where $\Im[\cdot]$ denotes the imaginary part of the complex argument, ρ_n is short for $\rho(\delta_n,\omega,\tau)$ and

$$\rho(\delta_n,\omega,\tau) = \frac{1}{2}\left|X_1(\omega,\tau)e^{-j\omega\delta_n} - X_2(\omega,\tau)\right|^2. \tag{8}$$

The original algorithm assumes two sensors, so a modification is made to make use of an arbitrary number of sensors in a linear array. Equation (6) is modified to

$$\delta_n(k) = \delta_n(k-1) - \mu\sum_{m=1}^{M-1}\frac{\partial J_{m,m+1}(\tau)}{\partial \delta_n} \tag{9}$$

where M is the number of sensors and $\partial J_{m,m+1}(\tau)/\partial\delta_n$ indicates the use of X_m and X_{m+1} instead of X_1 and X_2 in (7) and (8).

2.3. Demixing

The original algorithm performs the demixing using binary masks. The mask is defined as

$$\Omega_n(\omega,\tau) = \begin{cases}1 & \rho_n \le \rho_m, \qquad \forall m \ne n\\ 0 & \text{otherwise}\end{cases} \tag{10}$$

and the original source estimates in time-frequency representation is

$$\hat{S}_{n,m}(\omega,\tau) = \Omega_n(\omega,\tau)X_m(\omega,\tau) \tag{11}$$

where subscript m represents any one of the received mixtures.

At this point, the original algorithms reconstructs the original sources by transforming the time-frequency representation into time domain signals. This paper, however, will leave the sources in their time-frequency representation as the goal is not to reconstruct the signals but to identify the inter-sensor delay for each source, and the time-frequency representation is the needed representation for the delay estimation algorithm.

Furthermore, the demixing stage masks only a single mixture to create the time-frequency representation for the sources. It is necessary to further modify the original algorithm to mask all mixtures, as the delay estimation algorithm needs the separated sources from each sensor, not just a single sensor.

3. DELAY ESTIMATION

3.1. The Generalized cross correlation

The method used to estimate inter-sensor delays is based on the Generalized cross correlation method, described in [3]. The delay for source $n \in [1\ldots N]$ is estimated by maximizing the cross correlation between two signals $S_{m_1}(\omega,\tau)$ and $S_{m_2}(\omega,\tau)$, where $S_m = \hat{S}_{n,m}$ in (11), and can be expressed as

$$\hat{\Delta} = \arg\max_{\Delta} R_{S_{m_1}S_{m_2}}(\Delta). \tag{12}$$

The cross correlation $R_{S_{m_1}S_{m_2}}(\Delta)$ is related to the cross power spectrum $G_{S_{m_1}S_{m_2}}(\omega,\tau)$ by the Fourier transform as

$$R_{S_{m_1}S_{m_2}}(\Delta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_{S_{m_1}S_{m_2}}(\omega,\tau)\,e^{j\omega\Delta}d\omega. \tag{13}$$

The cross power spectrum can be calculated as

$$G_{S_{m_1}S_{m_2}}(\omega,\tau) = S_{m_1}(\omega,\tau)\,S_{m_2}^*(\omega,\tau) \tag{14}$$

where $(\cdot)^*$ denotes complex conjugate. The generalized cross correlation is defined in [3] as

$$R_{S_{m_1}S_{m_2}}(\Delta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi(\omega,\tau)\,G_{S_{m_1}S_{m_2}}(\omega,\tau)\,e^{j\omega\Delta}d\omega \tag{15}$$

where $\psi(\omega,\tau)$ is a general weighting function. The generalized correlation method known as the *phase transform*, or *PHAT*, is obtained by setting the weighting function to

$$\psi_{PHAT}(\omega,\tau) = \frac{1}{\left|G_{S_{m_1}S_{m_2}}(\omega,\tau)\right|}. \tag{16}$$

This weighting function normalizes the absolute value of all coefficients in the cross power spectrum to unity, and uses only the phase information to calculate the cross correlation. The PHAT weighting function have been found to work well in the presence of reverberation [5].

3.2. Multiple sensors

A generalization of the GCC-PHAT to handle multiple sensors is the SRP-PHAT algorithm and it is here defined as

$$\hat{\Delta} = \arg\max_{\Delta} \sum_{m_1=1}^{M-1} \sum_{m_2=m_1}^{M} \int_{-\infty}^{\infty} \frac{G_{S_{m_1} S_{m_2}}(\omega, \tau)}{\left| G_{S_{m_1} S_{m_2}}(\omega, \tau) \right|} e^{j\omega(m_2-m_1)\Delta} d\omega. \quad (17)$$

The original GCC-PHAT assumes two sensors, while SRP-PHAT generalizes into several sensors, where a delay is found that maximizes the cross correlation for all possible combinations of sensor pairs.

The SRP, or *steered response power*, principle is based on steering a beamformer across various locations searching for maximum output power. The beamformer is a delay-and-sum beamformer, which delays the output signals from the individual sensors and then sums them together to form the output.

3.3. Angle of arrival

When a delay is estimated, a corresponding angle of arrival can be calculated as

$$\hat{\alpha} = \arcsin\left(\frac{c \cdot \hat{\Delta}}{d}\right) \quad (18)$$

where c is the propagation speed of sound, $\hat{\tau}$ is the estimated time delay and d is the sensor separation distance. An angle of 0° corresponds to broadside, while -90° and +90° corresponds to the endfire directions.

Time delay estimation with a linear array is most accurate when the source is located near the broadside of the array, and the variance of the estimated angle will increase as the source approaches the endfire. The variance of the estimated angle is [6]

$$V[\hat{\alpha}] \propto \frac{V[\hat{\tau}]}{\cos^2 \alpha} \quad (19)$$

where $V[\cdot]$ denotes the variance operator and α is the true angle. If the source positions are restricted, the sensor array should be placed and oriented such that the source is located near the broadside as much as possible to keep the variance as low as possible.

Another issue is that the linear array can not determine if the source is in front of or behind the array. If only a two dimensional case is considered, positions that are mirrored along the line connecting the sensors results in the same relative time delays, which in turn will map to the same angle of arrival, even though the actual positions are different. This is, however, a limitation in the array geometry, and a different geometry can solve this problem. In this paper it is assumed that the source is restricted to only one side of the sensor array. In practice, this can be enforced by placing the array along a wall for example, effectively limiting the possible positions of the source.

4. FILTERING

In order to reduce the variance of the estimated angles, a filter is applied to the estimated values. The filter is a Kalman filter based on one-step prediction, as described in [7]. The Kalman filter is a state based filter, where the state vector contains all necessary information needed to predict future states assuming no external forces are acting on the system.

The state vector used in this paper only contains information about the current angle, but could also include information like rate of changes in the angle. Since the angle is a one dimensional quantity, the state vector at time index n is simply $\hat{\mathbf{x}}_n = [\alpha]$. The transition matrix $\mathbf{F}$ used to predict the state vector $\hat{\mathbf{x}}_{n+1}$ from $\hat{\mathbf{x}}_n$ is $\mathbf{F} = \mathbf{I}_1$, and the measurement matrix $\mathbf{C}$ used to extract the desired information from the state vector is $\mathbf{C} = \mathbf{I}_1$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. The correlation matrices for the process and measurement noise is $\mathbf{Q}_1 = q_1\mathbf{I}_1$ and $\mathbf{Q}_2 = q_2\mathbf{I}_1$, respectively, where q_1 and q_2 are the variances of the process and measurement noise.

```
for $n = 1, 2, 3\ldots$ do
    $\mathbf{G}_n = \mathbf{F} \cdot \mathbf{K}_n \cdot \mathbf{C}^H \cdot \left[\mathbf{C} \cdot \mathbf{K}_n \cdot \mathbf{C}^H + \mathbf{Q}_2\right]^{-1}$
    $\mathbf{a}_n = \mathbf{y}_n - \mathbf{C} \cdot \hat{\mathbf{x}}_n$
    $\hat{\mathbf{x}}_{n+1} = \mathbf{F} \cdot \hat{\mathbf{x}}_n + \mathbf{G}_n \cdot \mathbf{a}_n$
    $\mathbf{K}_{n+1} = \mathbf{F} \cdot \left[\mathbf{K}_n - \mathbf{F}^{-1} \cdot \mathbf{G}_n \cdot \mathbf{K}_n\right] \cdot \mathbf{F}^H + \mathbf{Q}_1$
end for
```

Table 1. Kalman filter based on one-step prediction.

The algorithm for estimating the state vector at iteration n, $\hat{\mathbf{x}}_{n+1}$, given estimated angles from the SRP-PHAT algorithm, $\mathbf{y}_n$, is shown in table 1.

An important feature of the state based model is that the state can be tracked for short periods even though the source is not active, since the state vector contains information to predict future states. In the context of speech localization, this can, for example, mean that the state vector is updated even during short pauses in the speech.

5. EVALUATION

The algorithms to estimate angle of arrivals for multiple concurrent speech sources is evaluated in a real room environment. The room represents a typical office room (hard walls, some furnitures etc.) of size $4 \times 5 \times 2{,}5$ meters. Speech sources are represented by loudspeakers, playing pre-recorded speech of random phrases. Figure 1 shows the four-microphone array and speaker setup. A speaker, representing the first source, is moved between the angles 0°, 20°, 40°and 60°. A second speaker, representing the second source, is placed at -30° throughout the test.

The tests focuses on measuring the variance and mean estimation error of the estimated angles after filtering as the first source moves between the four angles. The variance measures the deviation from the mean angle and indicates the amount of noise in the estimated angles, while mean estimation error is the addition of a static offset in the estimated angles compared to the real angle.

A Hanning window of 512 samples, with a 50% overlap, is used, and the sample rate is 16 kHz.

Top half of figure 2 shows the standard deviation. Two important properties are shown; as the angle for source 1 approaches the endfire of the array, the standard deviation increases, as implied in (19), and as the two sources are close to each other, the standard deviation also increases. When the sources are separated enough, the standard deviation of the second source remains constant.

Bottom half of figure 2 shows the mean estimation error. Again, when the two sources are separated enough, the mean estimation error for the second source remains constant. When the sources gets too close, they start to affect each other, implying there is a limit on how close two sources can be to be uniquely separated. The mean estimation error for the first source increases as the source approaches the endfire.

A second test is performed to evaluate how attenuated sources affect the variance of the angle estimates. The first source is placed at 40°, and the second at -30°, and the first source is attenuated.

Figure 3 shows the level of time-frequency orthogonality, i.e. non-overlapping area in time-frequency domain, calculated as in [4], for the two sources as the first source is attenuated. The figure also shows the standard deviation of the corresponding angle estimates as the first source is attenuated. When the first source is attenuated, the level of orthogonality decreases, but the standard deviation remains constant, which indicates robustness with respect to level differences of the sources.

6. CONCLUSIONS AND FUTURE WORK

Real room recordings have shown that the combination of algorithms in this paper forms a robust method for angle of arrival estimation for multiple concurrent speech sources. Good results were obtained in environments with moderate reverberation.

All steps involved in estimating the angles are suitable for real time applications, which is important as the system is intended for use with real time speech localization. The algorithms are also numerically simple enough to be performed in real time by a standard desktop computer.

In the problem of locating speech sources, this paper describes a method for estimating the angle of arrival from a single sensor array. The problem of finding the actual position still remains. By using multiple sensor arrays, a linear intersection algorithm as described in [8] can be used to determine intersection points for angle of arrivals from several sensor arrays, which have been found to work well for single source cases. In the case of multiple sources, it is necessary to match angle of arrivals from different sensor arrays such that the intersection points will correspond to actual sources. It is therefore necessary to investigate solutions to the problem of matching separated sources between sensor arrays.

7. REFERENCES

[1] Jean-François Cardoso, “Blind signal separation: statistical principles,” in *Proceedings of the IEEE, Special issue on blind identification and estimation*, 1998, vol. 9, pp. 2009–2025.

[2] Alexander Jourjine, Scott Rickard, and Özgür Yılmaz, “Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures,” *Proceedings ICASSP2000*, vol. 5, pp. 2985–2988, June 2000.

[3] Charles H. Knapp and G. Clifford Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, August 1976.

[4] Scott Rickard, Radu Balan, and Justinian Rosca, “Real-time time-frequency based blind source separation,” *Proceedings ICA2001*, pp. 651–656, December 2001.

[5] Anders Johansson, Nedelko Grbić, and Sven Nordholm, “Speaker localisation using the far-field srp-phat in conference telephony,” *ISPACS2002*, 2002.

[6] Steven M. Kay, *Fundamentals of statistical signal processing*, Prentice-Hall, 1993.

[7] Simon Haykin, *Adaptive filter theory*, Prentice Hall, fourth edition, 2002.

[8] Michael S. Brandstein, John E. Adcock, and Harvey F. Silverman, “A closed form location estimator for use with room environment microphone arrays,” *IEEE Transaction on Speech and Audio processing*, vol. 5, no. 1, pp. 45–50, January 1997.

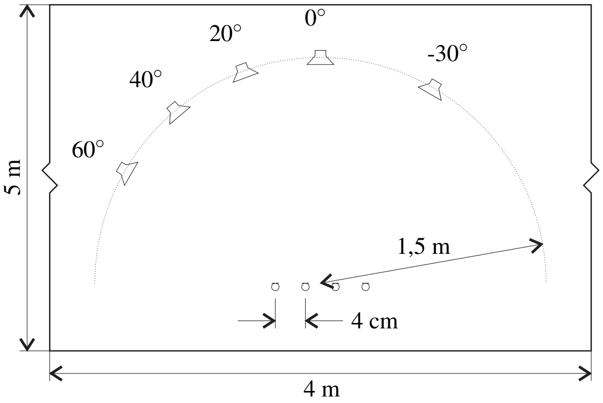

Fig. 1. Setup used to evaluate the performance of the algorithms. A loudspeaker is moved between the angles 0°, 20°, 40°and 60°, and a second loudspeaker is placed at -30°.

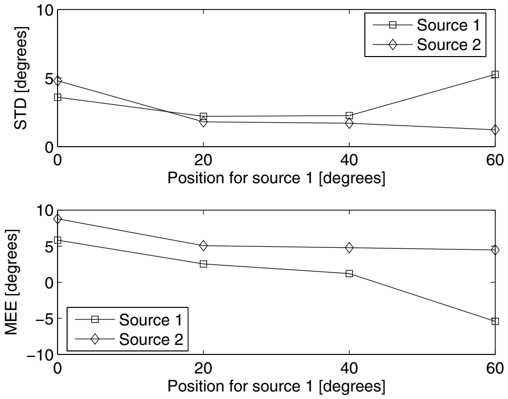

Fig. 2. Standard deviation (STD) and mean estimation error (MEE) for source 1 and source 2 as source 1 is moved to different angles.

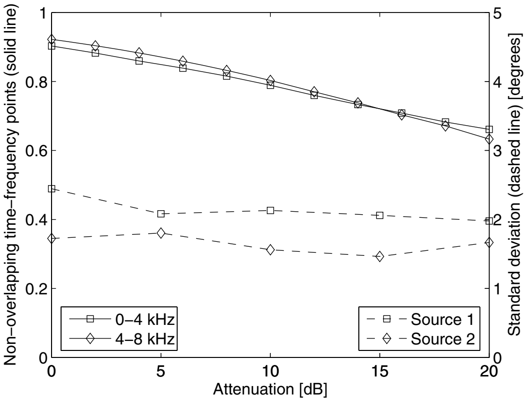

Fig. 3. Percentage of non-overlapping time-frequency points (solid line) and standard deviation (dashed line) with respect to level differences of two sources.