Multi-Source TDOA estimation using SNR-based angular spectra

Charles Blandin, Emmanuel Vincent, Alexey Ozerov

MULTI-SOURCE TDOA ESTIMATION USING SNR-BASED ANGULAR SPECTRA

Charles Blandin, Emmanuel Vincent and Alexey Ozerov

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France
{charles.blandin, emmanuel.vincent, alexey.ozerov}@inria.fr

ABSTRACT

This paper deals with the localization of multiple sources from two-channel mixtures recorded in a reverberant environment. We introduce new angular spectrum-based methods relying on the signal-to-noise ratio (SNR) to estimate the time difference of arrival (TDOA) of each source. We propose and compare five ways of estimating the SNR in each time-frequency point and in each direction, using beamforming techniques and statistical models. Large-scale evaluation considering a high number of situations shows the effectiveness of the proposed approach compared to state-of-the-art angular spectrum-based techniques.

Index Terms— Multiple source localization, TDOA estimation, signal-to-noise ratio, angular spectrum

1. INTRODUCTION

Recorded signals are often a mixture of several sound sources such as speech, music or noise. Source localization is the task of estimating the *direction of arrival* (DOA) of each source. It has potential applications in many domains such as videoconferencing, surveillance, or blind source separation.

This problem is particularly difficult in the two-channel under-determined case, when three or more sources must be localized from only two sensors. Localization is often achieved by finding *time difference of arrival* (TDOA) between channels for each source [5, 9]. Usually, this problem is addressed using the short-time Fourier transform (STFT). Let $\mathbf{X}(t,f) = [X_1(t,f), X_2(t,f)]^T$ and $S_n(t,f),\ n = 1,\ldots,N$ be respectively the STFT of the observed signals and the n-th source signal in time frame t and frequency bin f. The mixture can be modeled as the sum of a direct part and a reverberated part $\mathbf{B}(t,f)$:

$$\mathbf{X}(t,f) = \sum_{n=1}^{N} \mathbf{d}_{\tau_n}(f) S_n(t,f) + \mathbf{B}(t,f) \tag{1}$$

where τ_n is the TDOA of source n and

$$\mathbf{d}_\tau(f) = [1, e^{-2i\pi f\tau}]^T \tag{2}$$

is the *steering vector* associated with TDOA τ. Each TDOA thus translates into an expected phase difference $2\pi f\tau$ at each frequency f.

Three different approaches have been proposed in the literature. A first one [11, 3, 1] is to convert the observed phase difference into a TDOA in each time-frequency bin and to build a histogram of the TDOAs whose peaks point to the sources. This approach is restricted to small microphone spacings for which little or no spatial aliasing occurs. A second possibility [4, 8] is to alternately cluster the time-frequency bins into sources and update the source TDOAs according to the observed phase differences. It does not suffer from spatial aliasing but necessitates an initial guess of the source TDOAs due to local optima. A third approach [5, 9, 7] is to build a function of TDOA that we call angular spectrum, whose peak(s) indicate the TDOA(s) compatible with the observed phase difference in each time-frequency bin, and to sum this function over all bins. In the following, we consider the latter approach which is applicable to any microphone spacing and does not necessitate any prior guess of the source TDOAs.

A limitation of current angular spectrum-based methods is that they essentially assign the same weight to all observed phase differences, whether they result from the direct sound of single source or from a mixture of direct and reverberated sound and/or several sources[1]. In [1], a signal-to-noise ratio (SNR)-like confidence measure was proposed to weight the information provided in each time-frequency bin in the specific context of histogram-based localization in instantaneous mixtures and was shown to greatly improve localization performance.

In this paper, we propose to use the SNR in each time-frequency point to construct an angular spectrum and define five ways of estimating the SNR in a convolutive mixture using beamforming techniques and statistical models. In section 2, we describe the five proposed SNR-based angular spectra. We evaluate the performance of the proposed approaches and compare them to existing angular spectrum-based methods in section 3. Finally, we conclude in section 4.

This work was supported in part by French ANR project ECHANGE and by the Quaero Programme, funded by OSEO.

[1]MUSIC [9] attenuates the effect or reverberation or interfering sources by denoising in the parameter domain but the output spectrum does not depend on the resulting signal-to-noise ratio

2. SNR-BASED ANGULAR SPECTRA

In a given time-frequency bin (t, f), we define the $\mathrm{SNR}(t, f, \tau)$ associated with TDOA τ as the ratio between the signal power E_τ in this direction and the noise power E_b (power in all other directions). The SNR is then supposed to take large values in the direction of a given source in the time-frequency bins where this source is predominant. We propose to build an angular spectrum by summing the estimated SNRs over all time-frequency bins and all TDOAs:

$$\Sigma(\tau) = \sum_{(t,f)} \mathrm{SNR}(t, f, \tau) \tag{3}$$

The source TDOAs are then estimated by selecting the values of τ corresponding to the J highest peaks of $\Sigma(\tau)$. The choice of J leads to a trade-off between recall and precision , as described in section 3.

In order to estimate $\mathrm{SNR}(t, f, \tau)$, we propose three different approaches. The first one, that we call *a posteriori*, uses beamforming to estimate the source power and considers the residual power as noise. The second one, that we call *a priori*, jointly estimates the source and noise powers in the maximum likelihood (ML) sense, under a diffuse noise model. We then combine these two models to define a third approach based on frequency weighting of the *a posteriori* SNR.

2.1. Estimation of *a posteriori* SNR

From now on, we consider a single time-frequency bin (t, f) and omit its indices for simplicity. The power associated with TDOA τ can be estimated as the power of the output of a Delay-and-Sum (DS) beamformer or a Minimum Variance Distortionless Response (MVDR) beamformer, respectively given by [6]

$$E_\tau^{(\mathrm{DS})} = \frac{\mathbf{d}_\tau{}^H \widehat{\mathbf{\Phi}}_{\mathbf{xx}} \mathbf{d}_\tau}{4} \tag{4}$$

$$E_\tau^{(\mathrm{MVDR})} = (\mathbf{d}_\tau{}^H \widehat{\mathbf{\Phi}}_{\mathbf{xx}}^{-1} \mathbf{d}_\tau)^{-1} \tag{5}$$

where $\widehat{\mathbf{\Phi}}_{\mathbf{xx}}$ denotes the empirical mixture covariance matrix that can be computed as described in [2]. We write $E_b = E_X - E_\tau$ where $E_X = \frac{1}{2}\mathrm{tr}(\widehat{\mathbf{\Phi}}_{\mathbf{xx}})$ represent the total power. We obtain two ways of estimating the SNR for TDOA τ

$$\mathrm{SNR}_{\mathrm{DS}} = \frac{\mathbf{d}_\tau{}^H \widehat{\mathbf{\Phi}}_{\mathbf{xx}} \mathbf{d}_\tau}{2\mathrm{tr}(\widehat{\mathbf{\Phi}}_{\mathbf{xx}}) - \mathbf{d}_\tau{}^H \widehat{\mathbf{\Phi}}_{\mathbf{xx}} \mathbf{d}_\tau} \tag{6}$$

$$\mathrm{SNR}_{\mathrm{MVDR}} = \frac{(\mathbf{d}_\tau{}^H \widehat{\mathbf{\Phi}}_{\mathbf{xx}}^{-1} \mathbf{d}_\tau)^{-1}}{\frac{1}{2}\mathrm{tr}(\widehat{\mathbf{\Phi}}_{\mathbf{xx}}) - (\mathbf{d}_\tau{}^H \widehat{\mathbf{\Phi}}_{\mathbf{xx}}^{-1} \mathbf{d}_\tau)^{-1}} \tag{7}$$

Figures 1 (A) and (B) show the angular spectra obtained respectively with $\mathrm{SNR}_{\mathrm{DS}}$ and $\mathrm{SNR}_{\mathrm{MVDR}}$. As compared to the DS beamformer, MVDR beamformer appears to provide better noise elimination and enhances the peaks.

2.2. Estimation of *a priori* SNR by ML under a diffuse noise model.

In the *a posteriori* approach, the SNR is generally overestimated at low frequencies. Indeed, the observed phase differences are small so that $\mathbf{\Phi}_{\mathbf{xx}}$ becomes similar to $\mathbf{d}_\tau \mathbf{d}_\tau{}^T$ for all τ. This can be addressed by modeling both the source and the noise as random variables. We assume that in each time-frequency point, one source S of TDOA τ is predominant, and that S and noise $\mathbf{B}$ follow independent zero-mean Gaussian distributions. The mixture $\mathbf{X} = \mathbf{d}_\tau S + \mathbf{B}$ then follows a zero-mean Gaussian distribution with covariance matrix

$$\mathbf{\Phi}_{\mathbf{xx}} = v_s \mathbf{d}_\tau \mathbf{d}_\tau{}^H + v_b \mathbf{\Psi} \tag{8}$$

where v_s et v_b represent respectively the source variance and the noise variance, and $\mathbf{\Psi}$ is the covariance matrix of a diffuse noise [4, 2]

$$\mathbf{\Psi} = \begin{pmatrix} 1 & \mathrm{sinc}(2\pi f \frac{d}{c}) \\ \mathrm{sinc}(2\pi f \frac{d}{c}) & 1 \end{pmatrix} \tag{9}$$

where d is the distance between the two microphones, c is the soundspeed and $\mathrm{sinc}(.) = \frac{\sin(.)}{(.)}$. We estimate v_s and v_b in the ML sense using the closed form algorithm in [2]

$$\begin{pmatrix} v_s \\ v_b \end{pmatrix} = (\mathrm{diag}(\mathbf{\Lambda}_1) \;\; \mathrm{diag}(\mathbf{\Lambda}_2))^{-1} \, \mathrm{diag}(\mathbf{A}^{-1} \widehat{\mathbf{\Phi}}_{\mathbf{xx}} (\mathbf{A}^H)^{-1}) \tag{10}$$

where $\mathrm{diag}(.)$ denotes the column vector of diagonal entries of a matrix, $\mathbf{A}$ is the matrix whose columns are the eigenvectors of $\mathbf{d}_\tau \mathbf{d}_\tau{}^H \mathbf{\Psi}^{-1}$, and $\mathbf{\Lambda}_1$, $\mathbf{\Lambda}_2$ are equal respectively to $\mathbf{A}^{-1} \mathbf{d}_\tau \mathbf{d}_\tau{}^H (\mathbf{A}^H)^{-1}$ and $\mathbf{A}^{-1} \mathbf{\Psi} (\mathbf{A}^H)^{-1}$. Non-negativity is imposed by setting v_s to zero and v_b to $\frac{1}{2}\mathrm{tr}(\mathbf{\Psi}^{-1} \widehat{\mathbf{\Phi}}_{\mathbf{xx}})$ when v_b or v_s resulting from (10) is negative [2]. We then compute the SNR by:

$$\mathrm{SNR}_{\mathrm{APR}} = \frac{v_s}{v_b} \tag{11}$$

This method effectively addresses the SNR overestimation problem at low frequencies but results in slightly wider peaks, so that the resulting spectrum is not much different (see figure 1 (C)).

2.3. Estimation of *a priori* SNR by frequency weighting of *a posteriori* SNR

We now want to combine both approaches to obtain an angular spectrum with the global shape of *a priori* SNR and the precision of *a posteriori* SNR. In order to do so, we express the relationship between these two forms of SNR, in the simple case where the input signal consists in a single source of TDOA $\tau = 0$ and a diffuse noise. $\widehat{\mathbf{\Phi}}_{\mathbf{xx}}$ is then given by (8). By plugging (8) into (6) and (7), we obtain:

$$\mathrm{SNR}_{\mathrm{DS}} = \frac{1 + 2\,\mathrm{SNR}_{\mathrm{priori}} + \mathrm{sinc}(2\pi f \frac{d}{c})}{1 - \mathrm{sinc}(2\pi f \frac{d}{c})} \tag{12}$$

$$\text{SNR}_{\text{MVDR}} = \frac{1 + 2\,\text{SNR}_{\text{priori}}}{1 - \text{sinc}(2\pi f \frac{d}{c})} \tag{13}$$

where $\text{SNR}_{\text{priori}} = v_s/v_b$. By inverting these equations, we obtain a new way of computing *a priori* SNR by frequency weighting of *a posteriori* SNR, for both SNR_{DS} and SNR_{MVDR}:

$$\text{SNR}_{\text{DSW}} = W_d(f)\text{SNR}_{\text{DS}} + W_d(f) - 1 \tag{14}$$

$$\text{SNR}_{\text{MVDRW}} = W_d(f)\text{SNR}_{\text{MVDR}} - \frac{1}{2} \tag{15}$$

where $W_d(f) = \frac{1-\text{sinc}(2\pi f \frac{d}{c})}{2}$ is a frequency-dependent factor reducing the weight of low frequencies. Figure 1 (D) shows the shape of $W_d(f)$ for different values of d. Angular spectra obtained with SNR_{DSW} and $\text{SNR}_{\text{MVDRW}}$ are represented in figure 1 (E) and (F).

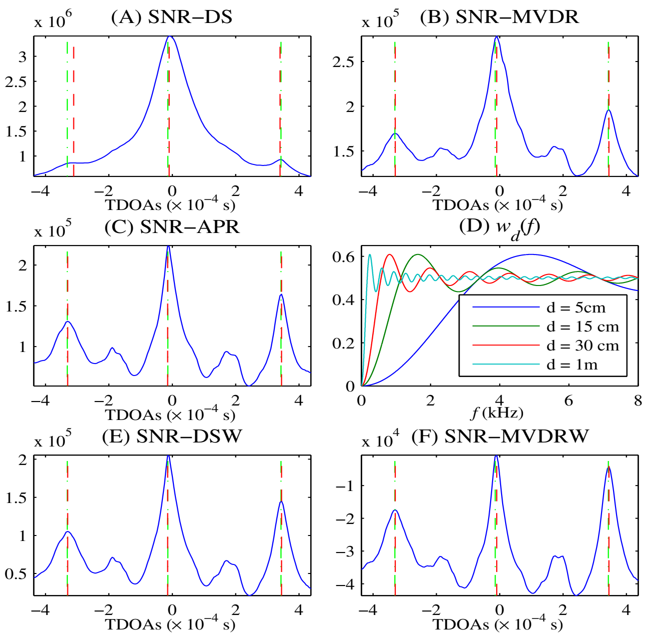

Fig. 1. Angular spectra produced by SNR_{DS} (A), SNR_{MVDR} (B), SNR_{APR} (C), SNR_{DSW} (E) and $\text{SNR}_{\text{MVDRW}}$ (F) for three female speech sources placed at 50 cm from the center of the microphone pair, with $d = 15$ cm and a reverberation time of 500 ms. (D) represents the weighting factors $W_d(f)$ for different microphone spacings d.

3. EXPERIMENTAL EVALUATION

We evaluated the five proposed methods on a large number of configurations, involving two to six sources, six reverberation times (from 50 ms to 750 ms), four microphone spacings (from 5 cm to 1 m), four distances between the sources and the center of the microphone pair (from 20 cm to 2 m), several source DOAs, and three source types (male speech, female speech and music). 4446 mixtures of 11 s duration were generated in total using impulse responses simulated via the Roomsimove toolbox[2] for a room of dimensions 4.45 m × 3.55 m × 2.5 m. Matlab code implementing the five proposed methods is available[3]. We compared these methods with three existing angular spectrum methods GCC-PHAT [5], MUSIC [9], and cSCT [7]. The parameters of cSCT were fixed by interpolation of the values chosen by the author for other configurations.

	SNR					GCC PHAT	MUSIC	cSCT
	DS	MVDR	APR	DSW	MVDRW			
$\mathcal{R}$	0.41	0.58	0.64	0.63	0.66	0.56	0.58	0.61
$\mathcal{P}$	0.48	0.61	0.66	0.67	0.69	0.65	0.24	0.64
$\mathcal{F}$	0.43	0.59	0.65	0.65	0.67	0.58	0.34	0.62

Table 1. Recall, precision and F-measure for $J = J_{\text{opt}}(\mathcal{A}, N, d)$ averaged over all configurations.

Evaluation was made in terms of recall, precision and F-measure. An estimated TDOA $\hat{\tau}$ is considered to be a correct estimate of a true TDOA τ if $\frac{c}{d}(\tau - \hat{\tau}) \leq \gamma$ with γ a constant set in our experiments to 0.05. For N sources, if we select the TDOAs corresponding to the J highest peaks of the angular spectrum, and we note I_J the number of correct TDOAs, we define the recall $\mathcal{R}$, the precision $\mathcal{P}$ and the F-measure $\mathcal{F}$ by $\mathcal{R}(J) = \frac{I_J}{N}$, $\mathcal{P}(J) = \frac{I_J}{J}$ and $\mathcal{F}(J) = 2\frac{\mathcal{R}(J)\times\mathcal{P}(J)}{\mathcal{R}(J)+\mathcal{P}(J)}$ respectively [10].

Figure 2 (A) shows the average F-measure as a function of J for $N = 6$ sources. For all evaluated algorithms, excepted MUSIC and GCC-PHAT, the F-measure reaches its maximum for J equal to the number of sources N. The best F-measure is obtained by $\text{SNR}_{\text{MVDRW}}$ and SNR_{DSW} which provide respectively an improvement of 0.06 and 0.04 compared to cSCT.

In the following, we fix the number of selected peaks to a value $J = J_{\text{opt}}(\mathcal{A}, N, d)$ for each algorithm $\mathcal{A}$, each number of sources N and each distance between microphones d, so as to maximise the F-measure averaged over all other parameters. Indeed, preliminary experiments showed that these parameters have the most important effect on the value of J_{opt}. The resulting average recall, precision and F-measure are presented in table 1. The best recall, precision and F-measure are obtained by $\text{SNR}_{\text{MVDRW}}$ and SNR_{DSW}.

Figure 2 shows the F-measure obtained as a function of microphone spacing d and reverberation time RT_{60} for speech sources. All methods provide poorer results with microphone spacings smaller than 15 cm or with larger reverberation time. $\text{SNR}_{\text{MVDRW}}$ outperforms the other evaluated methods for most microphone spacing and for all reverberation times. The two principal reasons are that it is robust to high reverberation thanks to MVDR beamforming, and to

[2]http://www.irisa.fr/metiss/members/evincent/software
[3]http://bass-db.gforge.inria.fr/bss_locate

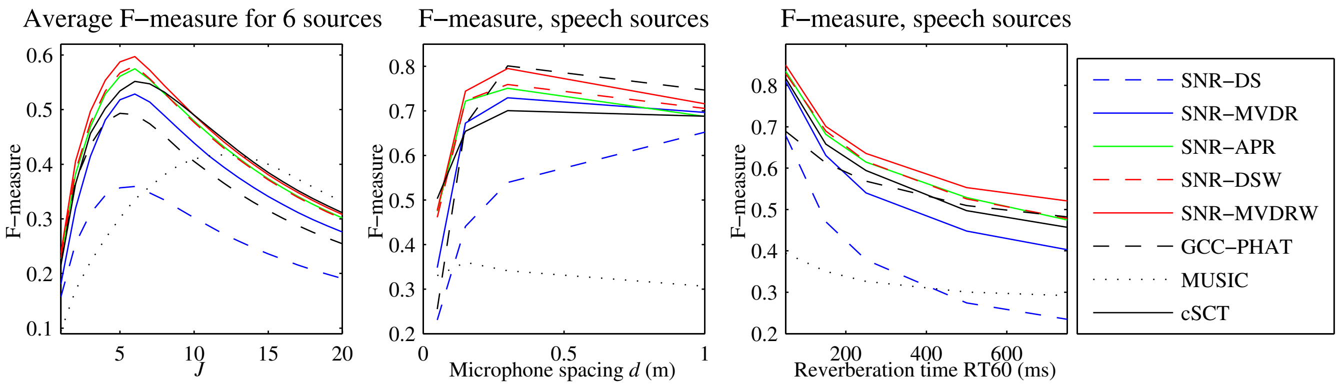

Fig. 2. Average F-measure of the evaluated algorithms as a function of the number of selected peaks J for all 6-source mixtures (A) and as a function of microphone spacing d (B) and reverberation time RT_{60} (C) with $J = J_{\mathrm{opt}}$ for all speech mixtures.

small microphone spacing thanks to frequency weighting.

4. CONCLUSION

We proposed five new angular spectrum methods using SNR to estimate TDOAs in two-channel under-determined mixtures. Large-scale evaluation showed that two of the proposed angular spectra, based on frequency weighting of SNR estimated by beamforming techniques outperform state-of-the-art methods in most configurations. Future work will focus on summing a nonlinear function of SNR in each time-frequency point and in using other information such as harmonicity to improve TDOA estimation.

5. ACKNOWLEDGEMENTS

The authors would like to thank F. Nesta for sharing his cSCT code, and N. Ito for discussion during the course of this work.

6. REFERENCES

[1] S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1):121–133, 2010.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval. Spatial covariance models for under-determined reverberant audio source separation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 129–132, 2009.

[3] C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116 (5):3075–3089, 2004.

[4] Y. Izumi, N. Ono, and S. Sagayama. Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 147–150, 2007.

[5] C. Knapp and G. Carter. The generalized cross-correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

[6] H. Krim and M. Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996.

[7] F. Nesta, P. Svaizer, and M. Omologo. Cumulative state coherence transform for a robust two-channel multiple source localization. In *Proc. 8th Int Conf on Independent Component Analysis and Signal Separation*, pages 290–297, 2009.

[8] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1592–1604, 2007.

[9] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[10] C.J. van Rijsbergen. *Information retrieval, 2nd Edition*. Butterworths, London, UK, 1979.

[11] O. Yilmaz and S.T. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.