Improving narrowband DOA estimation of sound sources using the complex Watson distribution

Anastasios Alexandridis and Athanasios Mouchtaris
FORTH-ICS, Heraklion, Crete, Greece, GR-70013
University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-70013
E-mail: {analexan, mouchtar}@ics.forth.gr

***Abstract*—Narrowband direction-of-arrival (DOA) estimates for each time-frequency (TF) point offer a parametric spatial modeling of the acoustic environment which is very commonly used in many applications, such as source separation, dereverberation, and spatial audio. However, irrespective of the narrowband DOA estimation method used, many TF-points suffer from erroneous estimates due to noise and reverberation. We propose a novel technique to yield more accurate DOA estimates in the TF-domain, through statistical modeling of each TF-point with a complex Watson distribution. Then, instead of using the microphone array signals at a given TF-point to estimate the DOA, the maximum likelihood estimate of the mode vector of the distribution is used as input to the DOA estimation method. This approach results in more accurate DOA estimates and thus more accurate modeling of the acoustic environment, while it can be used with any narrowband DOA estimation method and microphone array geometry.**

I. Introduction

Microphone arrays have received significant attention due to their superior performance over single microphones. They can spatially sample the sound field enabling a parametric spatial modeling of the acoustic environment in which typically multiple sound sources are active. This parametric approach finds numerous applications, such as signal enhancement, source separation, dereverberation, and spatial audio [1]. Typically, the parametric spatial modeling is achieved using instantaneous direction-of-arrival (DOA) estimates for each time-frequency (TF) point of the captured signals. Such modeling has been used in [2], [3] to design filters that extract the target speaker(s) from the recorded mixture. Instantaneous DOA estimates from multiple distributed microphone arrays which are fused together in order to estimate the instantaneous source locations are used in [4], [5]. Other applications include parametric spatial audio reproduction [6]–[8]—where instantaneous DOA estimates control the direction that each TF-point will be reproduced—and dereverberation [9].

To estimate the instantaneous DOAs, a narrowband DOA estimation method is applied in each TF-point. Of course, the performance of any method that utilizes this parametric spatial modeling depends on how accurately the instantaneous DOAs are estimated. Irrespective of the narrowband DOA estimation method used, some TF-points will suffer from erroneous DOA estimates which can occur due to noise and/or reverberation.

This research has been partly funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 644283, Project LISTEN.

In this paper, we propose a novel approach to more accurately estimate the instantaneous per TF-point DOAs. Currently, the narrowband DOA estimation is applied to the microphone signals at each TF-point. We propose an approach where instead of using the raw microphone signals in the TF-domain as input to the DOA estimation, an alternative input is generated through statistical modeling of the microphone signals at each TF-point. More specifically, we utilize the complex Watson distribution to model the microphone signals at a given TF-point and infer the maximum likelihood estimate of the distribution's mode vector, which is then used as input by the DOA estimation. The choice of the distribution is motivated by directional statistics where it is used to model uncertainties about directions of complex unit-norm vectors.

This distribution has been already used in audio signal processing, although in different ways: in [10]–[12] all TF-points are used together and form a mixture of complex Watson distributions, which is utilized for speech separation [10], while in [11], [12] variational inference on the mixture parameters is employed to determine the number of mixture components which corresponds to the number of active sources. Lastly, the complex Watson distribution is used in [13] as a distance metric for wideband DOA estimation of a single source.

In this paper, we use a different approach and utilize a complex Watson distribution with different parameters for each TF-point in order to provide an alternative input to a narrowband DOA estimation method. It is important to note that the outcome of our proposed method can be given as input to any narrowband DOA estimation method and thus our approach is independent of the DOA estimation method used and the array geometry. Our results, using simulations and real recordings, indicate that when the proposed method is used to generate the input for the DOA estimation, more accurate instantaneous DOA estimates and thus more accurate modeling of the acoustic environment can be achieved.

II. Problem Formulation

Consider a reverberant enclosure with P active sources and a microphone array with M microphones. The microphone signals in the Short-Time Fourier Transform (STFT) domain for time frame t and frequency bin f, are given by:

$$\boldsymbol{X}(t,f) = \sum_{p=1}^{P} \boldsymbol{H}_p(f) S_p(t,f) + \boldsymbol{N}(t,f) \qquad (1)$$

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

where $\boldsymbol{X}(t,f)$, $\boldsymbol{N}(t,f)$ are the $M \times 1$ vectors of microphone signals and noise, respectively, $S_p(t,f)$ denotes the signal of the p-th source, and $\boldsymbol{H}_p(f)$ is the $M \times 1$ vector of the frequency responses of the acoustic path from the p-th source to the microphones for frequency bin f.

Assuming that the speakers' signals are sparse and disjoint in the STFT domain [14], only one source p^* will dominate in each TF-point (t,f), thus the estimated DOA at this TF-point is expected to correspond to the direction of the source p^*.

III. Proposed methodology

We propose an approach to infer more accurate instantaneous per TF-point DOA estimates by changing the input given to the narrowband DOA estimation method. The microphone array signals are divided into frames of L samples with a time shift of K samples, windowed with a Hamming window. In the sequel, we omit the dependency on the time frame index t as the procedure is repeated for every frame.

Let $\boldsymbol{X}(f) = \left[X_1(f), \ \ldots, \ X_M(f)\right]^T$ be the $M \times 1$ vector of microphone signals in the frequency domain for frequency bin f, using an L-length Fourier Transform. The traditional approach (Fig. 1a) would utilize $\boldsymbol{X}(f)$ as input to the narrowband DOA estimation method in order to infer the DOA estimate $\theta(f)$. In our approach, we model the Fourier coefficients in $\boldsymbol{X}(f)$ with a complex Watson distribution. Then, to estimate $\theta(f)$ we use the maximum likelihood estimates of the distribution parameters as input to the DOA estimation method (Fig. 1b), instead of vector $\boldsymbol{X}(f)$ itself.

A. The complex Watson distribution

Let $\boldsymbol{y}$ denote a unit-norm d-dimensional complex random variable. The complex Watson distribution maps the observations of $\boldsymbol{y}$ to the d-dimensional complex unit hypersphere. The distribution is governed by the complex mode vector $\boldsymbol{\mu}$ and the real-valued concentration parameter κ, which describes how much the observations are concentrated around the mode vector. When $\kappa = 0$, the observations are uniformly distributed around the complex hypersphere. The probability density function of the complex Watson distribution is defined as [15]:

$$p(\boldsymbol{y}; \boldsymbol{\mu}, \kappa) = \frac{1}{c_{\mathcal{W}}(\kappa)} e^{\kappa |\boldsymbol{\mu}^H \boldsymbol{y}|^2} \tag{2}$$

where $(\cdot)^H$ denotes the Hermitian transpose operator and $c_{\mathcal{W}}(\kappa)$ is the normalizing constant which is given by:

$$c_{\mathcal{W}}(\kappa) = \frac{2\pi^d \mathcal{M}(1, d, \kappa)}{(d-1)!} \tag{3}$$

with $\mathcal{M}(\cdot)$ being Kummer's confluent hypergeometric function [16].

Given a set $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$ that contains n observation vectors from a complex Watson distribution, maximum likelihood estimates of the distribution's parameters can be found by forming the $d \times d$ matrix $\boldsymbol{\Phi}_{\boldsymbol{y}}$ as:

$$\boldsymbol{\Phi}_{\boldsymbol{y}} = \sum_{i=1}^{n} \boldsymbol{y}_i \boldsymbol{y}_i^H \tag{4}$$

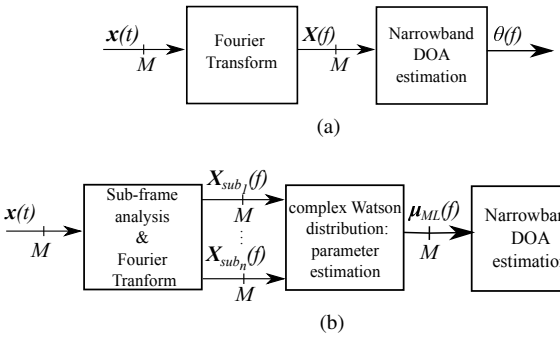

Fig. 1. Block diagram for the estimation of the instantaneous DOA for frequency point f at time frame t, using (a) the traditional input from the microphones and (b) our proposed methodology.

Let $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$ be the eigenvalues of $\boldsymbol{\Phi}_{\boldsymbol{y}}$ and $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ the corresponding eigenvectors. Then, the maximum likelihood estimate for the mode vector is given as the eigenvector that corresponds to the largest eigenvalue:

$$\boldsymbol{\mu}_{\text{ML}} = \boldsymbol{u}_1 \tag{5}$$

Although for the method presented in this paper the maximum likelihood estimate of the concentration parameter needs not be estimated, we report here for completeness that it can be found as the solution to the equation:

$$\frac{\frac{\partial}{\partial \kappa_{\text{ML}}} \mathcal{M}(1, d, \kappa_{\text{ML}})}{\mathcal{M}(1, d, \kappa_{\text{ML}})} = \frac{1}{N} \boldsymbol{\mu}_{\text{ML}}^H \boldsymbol{\Phi}_{\boldsymbol{y}} \boldsymbol{\mu}_{\text{ML}} \tag{6}$$

which is highly non-linear since it involves ratios of confluent hypergeometric functions. As a result, one has to resort to numerical approximations to estimate κ_{ML} [15].

B. Processing the microphone signals

Our processing starts by dividing an L-length frame into sub-frames of length $L_{\text{sub}} < L$ samples with a time shift of K_{sub} samples windowed with a Hamming window. The number of sub-frames is given by:

$$n = \lfloor L / K_{\text{sub}} \rfloor - 1 \tag{7}$$

where $\lfloor \cdot \rfloor$ denotes the floor operator.

The sub-frames are transformed into the frequency domain using an L-length Fourier Transform, resulting for each frequency $f = 1, \ldots, L$ in the set of Fourier coefficients

$$\mathcal{Y}(f) = \{\boldsymbol{X}_{\text{sub}_1}(f), \boldsymbol{X}_{\text{sub}_2}(f), \ldots, \boldsymbol{X}_{\text{sub}_n}(f)\} \tag{8}$$

where $\boldsymbol{X}_{\text{sub}_i}(f)$ is a $M \times 1$ vector of Fourier coefficients from the M microphones for the i-th sub-frame.

The samples in $\mathcal{Y}(f)$ form the observation vectors which are normalized to unit-norm according to:

$$\bar{\boldsymbol{X}}_{\text{sub}_i} = \boldsymbol{X}_{\text{sub}_i} / ||\boldsymbol{X}_{\text{sub}_i}|| \tag{9}$$

where $||\boldsymbol{x}|| = \sqrt{\boldsymbol{x}^H \boldsymbol{x}}$ denotes the norm of the vector $\boldsymbol{x}$. Unit-norm normalization represents a mapping of the observation vectors to the complex unit hypersphere, albeit preserving the

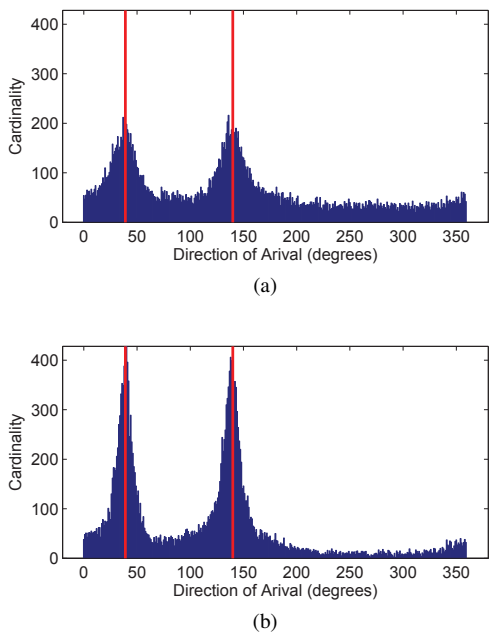

Fig. 2. Histogram of instantaneous DOA estimates obtained using MUSIC with (a) the traditional input and (b) the input derived by the proposed methodology, for a simulated recording of two active sources at 39° and 140° with reverberation time of $T_{60} = 400$ ms and 20 dB SNR. The vertical lines correspond to the true sources' DOAs.

spatial information which is essential for the DOA estimation. The n normalized observation vectors in:

$$\bar{\mathcal{Y}}(f) = \{\bar{\boldsymbol{X}}_{\text{sub}_1}(f), \bar{\boldsymbol{X}}_{\text{sub}_2}(f), \dots \bar{\boldsymbol{X}}_{\text{sub}_n}(f)\} \qquad (10)$$

can now be assumed to follow a complex Watson distribution with parameters $\boldsymbol{\mu}(f)$ and $\kappa(f)$. The maximum likelihood estimate $\boldsymbol{\mu}_{\text{ML}}(f)$ for the mode vector can then be found using the samples in $\bar{\mathcal{Y}}(f)$ as described in Section III-A. Finally, DOA estimation using an arbitrary narrowband DOA estimation method is applied using the estimated mode vector $\boldsymbol{\mu}_{\text{ML}}(f)$ as input. Note that, we only change the input to the narrowband DOA estimation method: instead of the traditional approach that utilizes the frequency domain coefficients from the microphone signals at each TF-point, we utilize these coefficients to infer an estimate of the mode vector of a complex Watson distribution and give this estimate as input to the DOA estimation. Fig 1 presents block diagrams that compare the traditional approach to the one proposed in this paper. Also, as we do not alter anything in the DOA estimation procedure, our proposed methodology can be used with any narrowband DOA estimation method and array geometry.

IV. Results and Discussion

To evaluate the performance of our proposed approach, we used simulations and real recorded signals. We considered two narrowband DOA estimation methods: the one described in [17] and the well-known narrowband Multiple Signal Classification (MUSIC) algorithm [18]. We used both DOA estimation methods to estimate the instantaneous DOAs at each TF-point and compare their performance when using the traditional input to the methods (Fig. 1a) and the one derived by our proposed methodology (Fig. 1b). For processing, we used frames of $L = 2048$ samples with 50% overlap. The FFT size was 2048. For our methodology, each frame was divided into sub-frames of $L_{\text{sub}} = 256$ samples with a time shift of $K_{\text{sub}} = 128$ samples which results in $n = 15$ sub-frames, i.e., $n = 15$ observation vectors for each frequency in order to perform the maximum likelihood estimation.

A. Simulation results

We used the Image-Source method [19] to simulate a room of dimensions of $6 \times 4 \times 3$ meters, characterized by reverberation time of $T_{60} = 400$ ms. We used a uniform circular microphone array with $M = 8$ microphones and a radius of 5 cm. The array was placed at the center of the room at 1 m height. For the given array geometry, the highest frequency of interest in order to avoid spatial aliasing is $f_{\text{max}} = 4$ kHz. Thus, in all our results we consider narrowband DOA estimation in all TF-points below f_{max}.

In each simulation, the sound sources were speech recordings of equal power and duration of 3 seconds sampled at 44.1 kHz. The signal-to-noise ratio (SNR) at each microphone was measured as the power of each source signal to the power of the noise. To simulate different SNR values we added white Gaussian noise at each microphone, uncorrelated with the source signals and the noise signals at the other microphones.

We considered scenarios of two and three simultaneously active sources. To more accurately measure the performance around the array, each scenario was repeated 50 times and the sources were located at random directions around the array with uniform probability and an angular separation between them of at least 30°. The sources were placed at 1 m height and their distance from the array was set to 1.5 m.

First, to qualitatively show the advantage of our proposed methodology, we consider two active sources at 39° and 140° and 20 dB SNR. Fig. 2 depicts the histogram of instantaneous DOA estimates using MUSIC with the traditional input (Fig. 2a) and the input derived by our proposed methodology (Fig. 2b). It is obvious that when the proposed input is given, MUSIC can more accurately estimate the instantaneous DOAs: the cardinality of DOA estimates very close to the true sources' DOAs is increased substantially, while erroneous estimates which are far away from the sources occur less often.

To compare the DOA estimation accuracy of the two aforementioned DOA estimation methods when using the traditional input from the microphone array and when using as input the one derived by our proposed methodology, we count the percentage of TF-points in which an accurate instantaneous DOA has been estimated. We consider that a DOA is accurate if its absolute error from a source's true DOA is less than 10°. Fig. 3 depicts the results for two and three simultaneous sources. It can be observed that our proposed methodology results in more accurate instantaneous DOA estimates compared to using the traditional input, for all SNR cases and number of active sources and for both DOA estimation methods that we consider. Especially at the higher SNR values, the number

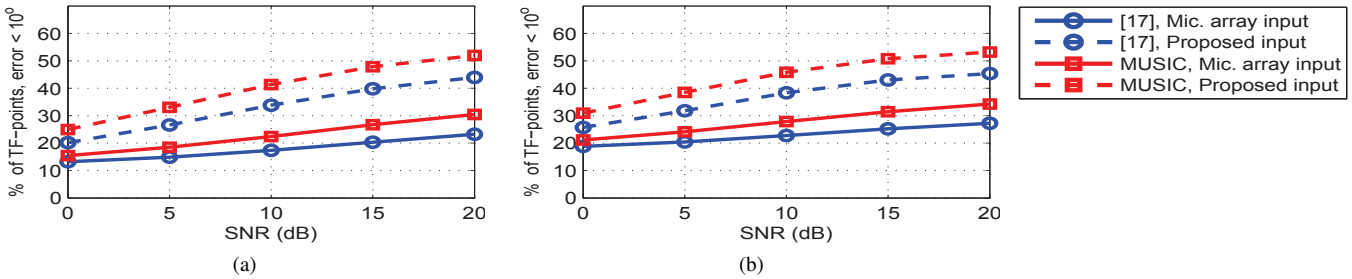

Fig. 3. Percentage of TF-points that exhibit DOA estimation error less than 10° for various SNR levels for the two DOA estimation methods when using the traditional input from the array and when using the input derived by our methodology for (a) two and (b) three simultaneously active sound sources.

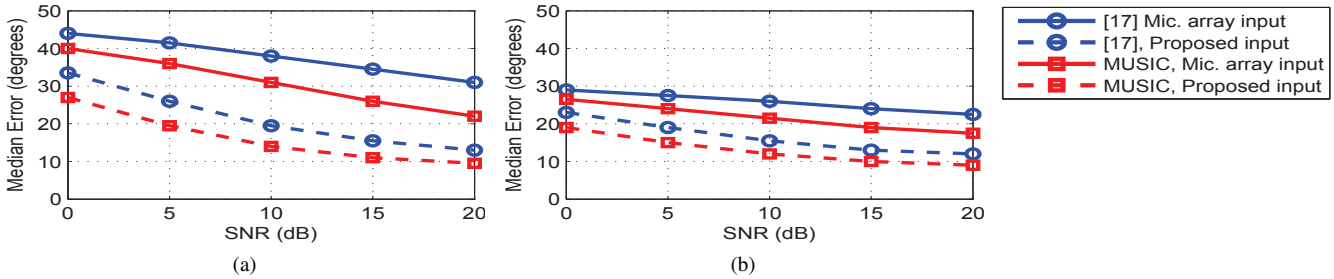

Fig. 4. Median DOA estimation error for various SNR levels for the two DOA estimation methods when using the traditional input from the array and when using the input derived by our methodology, for (a) two and (b) three simultaneously active sound sources.

of TF-points with accurate DOA estimates is increased by approximately 20% when using our proposed methodology, showing our method's capability to improve the accuracy of DOA estimation methods and to offer a more accurate parametric spatial modeling of the acoustic environment. Fig. 3 also reveals that, irrespective of the DOA estimation method, the traditional approach results in only a fraction of TF-points with accurate DOA estimates, while the majority of them suffer from noisy estimates. This highlights the need for more accurate approaches to estimate instantaneous DOAs and further motivates this study. Finally, as expected MUSIC performs better than the method of [17] either when the traditional or the proposed input to the method is used, which is due to the superior performance of subspace approaches to DOA estimation. Moreover, Fig. 4 depicts the corresponding median estimation error, which shows the gain in estimation accuracy that can be obtained when using our proposed methodology to generate the input for the DOA estimation method. It can be observed that our proposed methodology reduces the DOA estimation error by approximately 10° to 15° for all cases. Finally, comparing between the two and three source cases in Figs. 3 & 4, one can observe a better performance when three active sources are considered. Although, counter-intuitive at a first glance, it can be explained by the fact that we measure the DOA error of each TF-point from its closest source. Thus, some TF-points may exhibit reduced error, since it is more probable to find a source close to their estimated DOA when more sources are considered. However, this fact is of no significant importance, since our goal is to compare the two different inputs to the DOA estimation methods for the same scenarios.

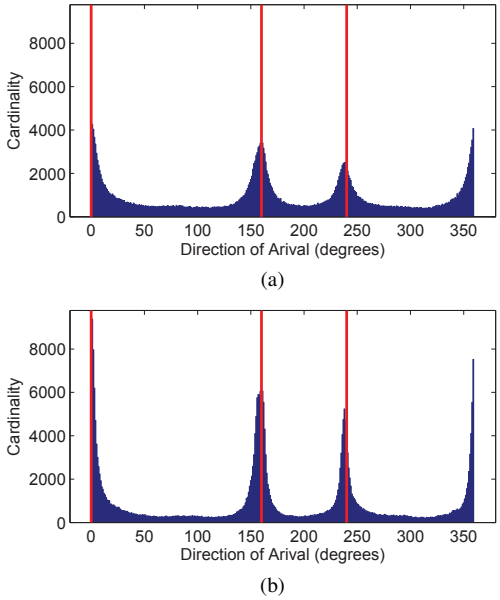

Fig. 5. Histogram of instantaneous DOA estimates obtained using MUSIC with (a) the traditional input and (b) the input derived by the proposed methodology, for a a real recording of three active sources at 0°, 160°, and 240° in a room of reverberation time of $T_{60} = 400$ ms. The vertical lines correspond to the true sources' DOAs.

B. Results using real recorded signals

We also conducted experiments in a typical office of approximately the same dimensions as in the simulations. We used a circular microphone array of 5 cm radius with eight Shure SM93 microphones and a TASCAM US2000 USB

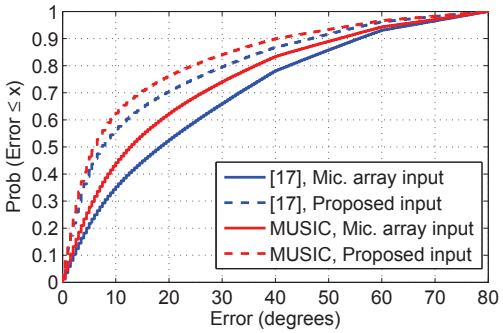

Fig. 6. Empirical Cumulative Distribution Function of the error of the instantaneous DOA estimates for the two DOA estimation methods when using the traditional input from the array and when using the one derived by our methodology, for a real recording with three active sound sources.

sound card. The reverberation time of the room was measured to be approximately $T_{60} = 400$ ms. We demonstrate the performance gain of our proposed methodology in a real recording of 45 seconds duration with three active speakers at $0°$, $160°$, and $240°$, and 1.5 m away from the array, which was placed on a table at the center of the room.

Fig 5 shows the histogram of instantaneous DOA estimates obtained with MUSIC with the traditional microphone array input (Fig. 5a) and the one obtained using the proposed methodology as input to MUSIC (Fig. 5b). In accordance with the simulations, it is evident that with our methodology MUSIC can infer much more accurate instantaneous DOA estimates: the cardinality of the DOAs estimated very close to the source's DOAs in much higher in Fig. 5b, while the cardinality of noisy DOA estimates that are away from the sources is reduced. Similar results were obtained in the histograms using the DOA estimation method of [17] which are omitted here due to space limitations. For each TF-point, we also calculated the absolute DOA estimation error for both DOA estimation methods with the traditional and proposed input. Fig. 6 depicts the empirical cumulative distribution function (CDF) of the error. The CDF shows the probability in the y-axis for the error to be less or equal to the corresponding value in the x-axis. The performance gain when using our proposed input to the DOA estimation methods is again evident. According to Fig. 6, in 50% of the cases (TF-points) the estimation error using the traditional input from the microphones is approximately 10 to 20 degrees depending on the method, and reduces to less than 6 degrees for both methods when our proposed input is considered. Finally, using the proposed input an approximately 20% gain is achieved in the number of TF-points whose DOA error is less than $10°$: from 35% when using the traditional input to 56% when using the proposed one for the method of [17], and from 44% to 62% for MUSIC.

V. Conclusions

In this work, we considered the parametric modeling of the acoustic environment with instantaneous DOA estimates for each TF-point. We argued that the traditional approach where the microphone signals in each TF-point are given as input to the narrowband DOA estimation method can result in noisy and inaccurate instantaneous DOAs, which was validated from our experiments. We proposed a novel approach where the input given to the DOA estimation method is derived through statistical modeling of each TF-point with a complex Watson distribution. As our proposed approach only alters the input to the DOA estimation method—and not the method itself—it can be applied to any DOA estimation approach and array geometry. Through simulations and experiments with real recorded signals, we showed that our proposed methodology achieves a significant gain in the instantaneous DOA estimation accuracy.

References

[1] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Sig. Proces. Mag.*, vol. 32, no. 2, pp. 31–42, March 2015.

[2] O. Thiergart, M. Taseska, and E. Habets, "An informed MMSE filter based on multiple instantaneous direction-of-arrival estimates," in *EUSIPCO*, Sept 2013, pp. 1–5.

[3] ——, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proces.*, vol. 22, no. 12, pp. 2182–2196, Dec 2014.

[4] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation and counting in a wireless acoustic sensor network," in *IEEE WASPAA*, Oct 2015, pp. 1–5.

[5] M. Taseska and E. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Trans. on Audio, Speech, and Lang. Proces.*, vol. 22, no. 7, pp. 1195–1207, July 2014.

[6] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, 2007.

[7] F. Kuech, M. Kallinger, R. Schultz-Amling, G. Del Galdo, J. Ahonen, and V. Pulkki, "Directional audio coding using planar microphone arrays," in *HSCMA, 2008.*, May 2008, pp. 37–40.

[8] M. Cobos, J. J. Lopez, and S. Spors, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 2:1–2:13, 2010.

[9] M. Kallinger, G. Del Galdo, F. Kuech, and O. Thiergart, "Dereverberation in the spatial audio coding domain," in *Audio Engineering Society Convention 130*, May 2011.

[10] D. H. T. Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *IEEE ICASSP*, March 2010, pp. 241–244.

[11] L. Drude, A. Chinaev, D. H. T. Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational EM approach for complex Watson mixture models," in *IEEE ICASSP*, May 2014, pp. 6834–6838.

[12] ——, "Towards online source counting in speech mixtures applying a variational EM for complex watson mixture models," in *IWAENC*, Sept 2014, pp. 213–217.

[13] L. Drude, F. Jacob, and R. Haeb-Umbach, "DOA-estimation based on a complex watson kernel method," in *EUSIPCO*, 2015, pp. 255–259.

[14] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE ICASSP*, vol. 1, May 2002, pp. 529–532.

[15] K. V. Mardia and I. L. Dryden, "The complex watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[16] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST Handbook of Mathematical Functions*, 1st ed. New York, NY, USA: Cambridge University Press, 2010.

[17] A. Karbasi and A. Sugiyama, "A new DOA estimation method using a circular microphone array," in *EUSIPCO*, 2007, pp. 778–782.

[18] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, Mar 1986.

[19] E. Lehmann and A. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Trans. on Audio, Speech, and Lang. Proces.*, vol. 18, no. 6, Aug. 2010.