REAL-TIME MULTIPLE DOA ESTIMATION OF SPEECH SOURCES IN WIRELESS ACOUSTIC SENSOR NETWORKS

David Ayllón[1]*, Roberto Gil-Pita*[1]*, Manuel Rosa-Zurera*[1] *and Hamid Krim*[2]

[1]Department of Signal Theory and Communications, University of Alcala, Spain
[2]Department of Electrical and Computer Engineering, North Carolina State University, NC

ABSTRACT

Indoor localization of multiple speech sources in wireless acoustic sensor networks (WASNs) is an open and interesting problem with many practical applications, but the presence of noise and reverberations complicates the problem. In this paper, a distributed algorithm for multiple DOA estimation of speech sources in WASNs is presented. The method exploits the sparsity of speech sources in the time-frequency domain to obtain DOA estimations locally in each node of the network. The DOA estimations of different nodes are further combined to increase the accuracy of the local DOA estimations. Since the local DOAs are estimated using only the microphones of the same node, the synchronization between input channels and localization of the microphones from different nodes are not an issue.

Index Terms— Speech source localization, wireless acoustic sensor networks, speech processing.

1. INTRODUCTION

The localization of multiple speech sources in closed environments has many trending applications, such as hands-free speech communications, videoconference systems, automatic surveillance systems or video games [1]. Additionally, some other speech-based applications require source localization as an intermediate step, for instance, speech enhancement, automatic speech recognition or speaker identification. Unfortunately, the ability to localize speech sources in a closed environment is highly affected by background noise and reverberations.

When the signals are observed by an array of sensors, the localization problem is addressed by estimating the time-difference-of-arrival (TDOA) of each source in each pair of sensors. A review of methods for multi-source TDOA estimation in reverberant audio can be found in [2]. TDOA estimation relies on the assumptions that the microphone positions (or at least their relative distances) are known and that the different input channels are synchronized. These two assumptions are usually met in a fixed microphone array but they are not so obvious in a wireless sensor network where the position of each node may be unknown by the other nodes in the network and there is a high probability that the input channels from different nodes are not synchronized. For instance, the works in [3, 4] propose a maximum likelihood (ML) acoustic source location estimation in a wireless sensor network. The first one assumes to know the position of each sensor and the second one assumes to know the position of some sources to pre-localize the nodes, so they can be considered semi-blind source localization methods.

This paper presents a novel method for blind localization of multiple speech sources in a wireless acoustic sensor network (WASN). Direction-of-arrival (DOA) estimations are obtained locally in each node, using a method based on the algorithm proposed in [5] to identify the number of speech sources in a mixture. This method uses a parametric model to estimate the probability density function (PDF) of time-frequency domain TDOA estimates, and its potential to estimate DOAs is investigated in this work. The localization algorithm assumes that each node has two microphones, although it is easily extensible to a higher number of microphones. The accuracy in the DOA estimation depends on the relative position between the source and the array, which causes that DOA estimations from different nodes have different accuracy. In order to increase the robustness of local estimations, a best linear unbiased estimator (BLUE) that combines the DOA estimations obtained in different nodes is also derived. Since the final DOA estimations are calculated from DOA estimations obtained locally in each node, the knowledge of the microphone positions and the synchronization between input signals is guaranteed. Additionally, the information from different nodes is considered and the computational load of the algorithm is distributed between them.

2. PARAMETRIC MODEL-BASED ESTIMATION OF TDOA IN THE TIME-FREQUENCY DOMAIN

Let us consider a microphone array j composed of $M = 2$ microphones and set N different sources, $s_n(t)$, $n \in \{1, \cdots, N\}$, impinging the array. The microphone signals are described by

This work has been funded by the Spanish Ministry of Economy and Competitiveness, under project TEC2012-38142-C04-02, and the University of Alcalá, under project CCG2013/EXP-076.

$$y_{jm}(t) = \sum_{n=1}^{N} h_{jmn}(t) * s_n(t) + n_{jm}(t), \quad m = 1, 2, \quad (1)$$

where the filter $h_{jmn}(t)$, commonly denominated acoustic impulse response, describes the acoustic channel between the n-th source and the m-th microphone of the j-th array, $n_{jm}(t)$ represents additive noise at the m-th sensor of the j-th array, and the operator $*$ represents linear convolution.

The short-time Fourier transform (STFT) of the microphone signals is represented by $Y_{jm}(k,l)$, where $k = 1, \cdots, K$ represents frequency, and $l = 1, \cdots, L$ the time frame. Assuming that the sources do not overlap in the TF domain, according to [6], an estimate of the TDOA in each TF point (TF-TDOA) can be obtained from the STFT of the signals of the two microphones:

$$\hat{\delta}_j(k,l) = -\frac{1}{\omega} arg\left\{\frac{Y_{j2}(k,l)}{Y_{j1}(k,l)}\right\}, \quad (2)$$

where ω represents angular frequency. Ideally, if the sources were completely separated in the TF domain, the TF-TDOA estimates $\hat{\delta}_j(k,l)$ would take the value of the true TDOA of the active source of each TF bin. However, this assumption is only approximated, and speech sources show some overlap in the TF domain. In this case, the TF-TDOA estimates $\hat{\delta}_j(k,l)$ will cluster around the true TDOA values of the different sources. In order to avoid phase ambiguity due to large microphone distances in the computation of $\hat{\delta}_j(k,l)$, the phase is previously unwrapped for all frequencies of each frame (changing absolute phase jumps greater or equal to π to their 2π complement).

The proposed method for TDOA estimation relies in the fact that, since the TDOA calculated in a pair of microphones changes with the position of the sources, the PDF of the TF-TDOA estimates has different modes associated with each of the sources. According to this, an estimate of the TDOA of each source can be obtained by finding the centroid of each of these modes. The method in [5] proposes to estimate the PDF of the TF-TDOA estimates using a parametric model-based method. Let us define the vector $\mathbf{d} = [\delta_1, ..., \delta_Q]$ containing all the TF-TDOA estimates from a pair of microphones, $\hat{\delta}_j$, where $Q = K \cdot L$. The PDF of the random variable δ is denoted by $f(\delta)$ and it is going to be modeled using an autoregresive (AR) model, which parameters a_p can be obtained with the Yule-Walker equations:

$$\hat{\phi}_\delta[m] = \sum_{p=1}^{P} a_p \hat{\phi}_\delta[m-p] + \sigma_\epsilon^2 \delta_K[m], \quad 0 \leq m \leq P, \quad (3)$$

where σ_ϵ^2 is the variance of the prediction error, and $\delta_K[m]$ is the Kronecker delta, and P is the order of the model. The values $\phi_\delta[m]$ can be obtained from the PDF, making use of the dual function (characteristic function) given by the next expression:

$$\phi_\delta[m] = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\delta m} f(\delta) d\delta = \frac{1}{2\pi} E\{e^{i\delta m}\}. \quad (4)$$

Using the sample mean as an estimator of the probabilistic expectation, the sequence $\phi_\delta[m]$ can be estimated by

$$\hat{\phi}_\delta[m] = \frac{1}{2\pi Q}\sum_{q=1}^{Q} e^{i\delta_q m}. \quad (5)$$

Finally, using the AR model, the PDF is estimated as

$$\hat{f}(\delta) = \frac{\sigma_\epsilon}{|1 - \sum_{p=1}^{P} a_p e^{-i2\pi\delta}|^2}. \quad (6)$$

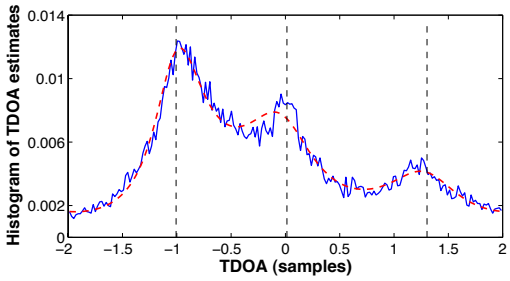

Fig. 1. Histogram-based estimate (dashed blue line) and 3th-order parametric-model estimate (solid red line) of the PDF of TF-domain TDOA estimate for a mixture of 3 speech sources with additive background noise of SNR=10 dB and reverberation time of 262 ms. Vertical lines represent the true TDOAs.

Figure 1 shows the estimate of the PDF of the TF-TDOA estimates, in the case of a mixture of 3 speech sources with additive background noise of SNR=10 dB and reverberation time of 262 ms, estimated with a histogram (dashed blue line) and estimated with the proposed parametric-model with order 3 (solid red line). The dashed black lines (vertical lines) represent the true TDOA of each source. Due to noise and reverberation, the width of the peaks that appear in the PDF is larger and they are not so well defined in comparison to a non-reverberant noiseless case. However, the smooth estimation of the PDF performed by the parametric linear model overcomes this problem, and it is clear how the method estimates the PDF with accuracy and the peaks obtained correspond with the true TDOA.

3. REAL-TIME MULTIPLE DOA ESTIMATION IN A WIRELESS ACOUSTIC SENSOR NETWORK

3.1. Local DOA estimates

Let us consider a general WASN with J nodes where, without loss of generality, each node j has access to two microphone signals denoted in the frequency domain as $Y_{j1}(k,l)$

and $Y_{j2}(k,l)$. According to this, the total number of microphones in the network is $M = 2J$. The application of the parametric model-based estimation of the PDF described in the previous section to estimate multiple TDOA in each node j is straightforward: the values of δ where the peaks appear in the PDF estimation correspond to the phase of the roots of the denominator polynomial of the AR model (expression (6)). Henceforth, $\hat{\delta}_n^j$ denotes the TDOA estimate for the n-th source in the j node and $\hat{\theta}_n^j$ its corresponding DOA.

In the practical implementation the algorithm is realized in the short-term spectral domain, obtaining DOA estimates in each time frame, which allows an implementation of the algorithm in real time. Time frames with power lower than the estimated background noise power are discarded. The power is calculated as the average power of the two microphones and the background noise power is estimated during periods of silence. First, TF-TDOA estimates in each frequency band $\hat{\delta}_j(k,l)$ are calculated using expression (2). Second, the sequence $\phi_\delta[m]$ is recursively estimated by means of the short-time estimates according to

$$\phi_\delta[m,l] = \alpha\phi_\delta[m,l-1] + (1-\alpha)\frac{1}{2\pi K}\sum_{k=1}^{K} e^{i\hat{\delta}_j(k,l)m}, \quad (7)$$

where α is a smoothing factor $0 < \alpha < 1$. This smoothing factor avoid sudden changes in the DOA estimations that may be caused by wrong estimation in some time periods. The AR model is fitted to the short-time sequence in (7) and TDOA estimates are obtained in each time frame from the calculated AR model. Finally, the estimation of the DOA of the n-th source in the j node and in the l-th time frame, which is denoted as $\hat{\theta}_n^j(l)$, is given by

$$\hat{\theta}_n^j(l) = cos^{-1}\frac{\hat{\delta}_n^j(l)c}{d_{mic}f_s}, \quad (8)$$

where $\hat{\delta}_n^j(l)$ is the TDOA estimate of the n-th source in the j node and in the l-th time frame, d_{mic} is the distance between the two microphones of the node, f_s is the sampling rate and c the speech of sound.

3.2. Distributed DOA estimation

Let us assume that each node of the network provides local DOA estimates in each time frame, thus having J different estimations of the DOA of each source. The next problem to solve is how to combine the J DOA estimations $\hat{\theta}_n^j(l)$ in such way that their combination improves the accuracy of the local estimations.

The DOA estimations obtained in each node are referred to its local Cartesian coordinate system, with the origin at the common midpoint of the two sensors. In order to combine the DOA estimations of different nodes, they should be first expressed in terms of a global Cartesian coordinate system via the appropriate translation and rotation. To do this operation, we assume that the distances between the origin of the local Cartesian coordinate systems of the different nodes are available (i.e. straight distance between nodes). The estimation of these distances can be made either acoustically or taking advantages of the wireless links available in the nodes, but the solution of this problem is out of the scope of this paper.

The local TDOA estimates $\hat{\delta}_n^j(l)$ are assumed to be independent (from node to node) and corrupted by Gaussian white noise with variance σ_j^2. This assumption has been confirmed experimentally obtaining a high number of realizations of the same experiment and in different acoustic scenarios. According to this, the best linear unbiased estimator (BLUE) [7] is given by

$$\hat{\theta}_n(l) = \frac{\sum_{j=1}^{J}\frac{\hat{\theta}_n^j(l)}{\sigma_j^2}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2}}. \quad (9)$$

The BLUE estimator weights those local DOA estimations with smaller variances more heavily than the ones with higher variances. The variance of the estimator σ_j^2 is directly related to the variance of the prediction error of the AR model used to obtain the TDOA estimates, σ_ϵ^2. Since the denominator of expression (9) make the estimator unbiased, the variance of the prediction error is used instead.

The complete steps of the proposed algorithm in the j-th node are the next.

1. Transform the two input signals into the frequency domain, $Y_{j1}(k,l)$ and $Y_{j2}(k,l)$.

2. Compare the energy of the current time frame against the threshold to decide whether it should be processed or not.

3. Calculate the TF-TDOA estimates, $\hat{\delta}_j(k,l)$ using expression (2) and update the sequence in (7).

4. Calculate the linear prediction coefficients and the variance of the prediction error using the Yule-Walker equations.

5. Find the phases of the roots of the polynomial in the denominator of the AR model, which are $\hat{\delta}_n^j(l)$, and calculate $\hat{\theta}_n^j(l)$ using expression (8).

6. Broadcast the local DOA estimates $\hat{\theta}_n^j(l)$ and the corresponding variances through the sensor network.

7. Receive the DOA estimates from other nodes, $\hat{\theta}_n^i(l), i = 1, \cdots J, i \neq j$.

8. Obtain the collaborative DOA estimations $\hat{\theta}_n(l)$ using expression (9).

4. SIMULATIONS

The accuracy in the DOA estimations of the proposed algorithm has been evaluated in the case of 1, 2 and 3 speech sources, in a simulated (8m x 7m x 3m) rectangular room with plane reflective surfaces and uniform, frequency-independent reflection coefficients. Room impulse responses were generated with the image method [8], using reverberation times T_{60} that range from 0 to 372 ms. The additive background noise is speech-shaped noise with SNR of 10 dB. The algorithm assumes to know the number of sources in advance (they can be estimated using [5]). The nodes have been placed in the center of the room, with a rectangular arrangement of 2x3m. Different source positions have been evaluated, varying de DOAs: 5 different scenarios in the case of one source, 4 scenarios in the case of 2 sources, and 3 scenarios in the case of 3 sources. In any case, the absolute difference between DOAs of any pair of sources is not smaller than 10°. Microphones and sources are in the same elevation plane. For each scenario, 50 different speech mixtures of 4 seconds length have been generated, using different speech sources, randomly selected from the TIMIT database [9]. The source signals have been normalized with equal power before mixing. The sampling rate is 44.1 kHz and the TF decomposition is performed by a STFT with frame length of 2048 samples, using a hamming window with 50% overlap. The accuracy of the localization algorithm has been measured by the frame average absolute error in the DOA estimation, averaged for N sources:

$$E_{DOA} = \frac{1}{NL} \sum_{n=1}^{N} \sum_{l=1}^{L} |\theta_n(l) - \hat{\theta}_n(l)|, \qquad (10)$$

where $\theta_n(l)$ is the true DOA for the n-th source in the l-th frame.

Table 1 contains the average E_{DOA} values for DOA estimations in the different simulated scenarios. It contains the average values for the different scenarios and mixtures as well as the standard deviations. In Case A, the mixtures were generated without noise and reverberations ($T_{60} = 0$) and, in Case B, the mixtures were generated with a T_{60}=263 ms and background noise with a SNR=10 dB. The results show that in Case A, the DOA estimations are very accurate for any number of sources, with a maximum mean error of 1.42 degrees in the case of N=3. The introduction of background noise and reverberations obviously decreases the accuracy: in the case of localizing a single source the average error is 2.61 degrees, in the case of two sources is 4.69 degrees, and in the case of four sources is 5.43 degrees. These results are good considering the noise level and reverberation. Additionally, the graph in 2 shows the effects of reverberation in the accuracy of the DOA estimations. The average values of E_{DOA} are represented for different levels of reverberation. The estimation error clearly increases with reverberation (almost monotonically). For instance, in the case of 2 sources, the average error is slightly higher than 1 degree without reverberation, but it is increased to values higher than 5 degrees for reverberation times of T_{60}= 372. Nevertheless, the estimation error is acceptable even for high levels of reverberation.

Table 1. Mean and standard deviation (degrees) of the average error E_{DOA} in the DOA estimation of N sources.

Accuracy	N=1		N= 2		N=3	
	Mean	STD	Mean	STD	Mean	STD
Case A	0.75	0.44	1.21	1.01	1.42	0.93
Case B	2.61	1.02	4.69	2.76	5.43	2.21

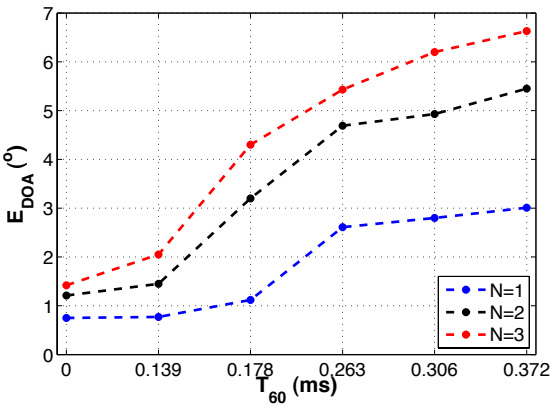

Fig. 2. Average accuracy for different levels of reverberation and different number of sources.

5. DISCUSSION

This work presents a novel method for multiple DOA estimation of speech sources in WASNs. Local DOA estimations are obtained frame by frame in each node using only two microphones, with an algorithm that has shown robustness agains noise and reverberation. The local DOA estimations from different nodes are combined to increase the accuracy of the estimation. The problems of synchronization between input channels and estimation of the microphones positions, which are common problems in WASN, are solved with the proposed schema. The computational load of the algorithm is distributed among the different nodes.

The presented results are promising, but further research is necessary in order to investigate the robustness of the method in real environments, the dependence of the performance on the relative positions between nodes and sources (testing more scenarios), and the estimation of the distance between nodes. Finally, although the results presented in this paper are not compared with other existing method for robust DOA estimation of multiple sources, the authors believe that the performance of the proposed method is noteworthy. This comparison, as well as further experiments required for generalization of the method, will be presented in a future extension of this work.

6. REFERENCES

[1] J. C. Chen, K. Yao, and R. E. Hudson, "Source localization and beamforming", *Signal Process. Magazine*, vol.19(2), pp. 30-39, 2002.

[2] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering, " *Signal Process.*, vol. 92, pp. 1950-1960, 2012.

[3] X. Sheng and Y.H. Hu, "Maximum Likelihood Multiple-Source Localization Using Acoustic Energy Measurements with Wireless Sensor Networks", *IEEE Trans. Signal Process.*, vol. 53, pp. 44-53, Jan 2005.

[4] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-Likelihood Source Localization and Unknown Sensor Location Estimation for Wideband Signals in the Near-Field, " *IEEE Trans. Signal Process.*, vol. 50(8) pp. 1843-1854, Aug 2002.

[5] D. Ayllon, R. Gil-Pita, M. Rosa-Zurera, and H. Krim, "An information theoretic approach for speech source enumeration," in *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, pp. 4300-4304, May 2013.

[6] O.Yilmaz and S.Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, pp. 1830-1847, July 2004.

[7] S. Kay, "Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory," vol. 1, Prentice Hall, 1993.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65(4), pp. 943-950, 1979.

[9] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," In *Proc. of the DARPA Workshop on speech recognit.*, pp. 93-99, 1986.