SOUND-MODEL-BASED ACOUSTIC SOURCE LOCALIZATION USING DISTRIBUTED MICROPHONE ARRAYS

Rupayan Chakraborty and Climent Nadeu

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona, Spain
{rupayan.chakraborty, climent.nadeu}@upc.edu

ABSTRACT

Acoustic source localization and sound recognition are common acoustic scene analysis tasks that are usually considered separately. In this paper, a new source localization technique is proposed that works jointly with an acoustic event detection system. Given the identities and the end-points of simultaneous sounds, the proposed technique uses the statistical models of those sounds to compute a likelihood score for each model and for each signal at the output of a set of null-steering beamformers per microphone array. Those scores are subsequently combined to find the MAP-optimal event source positions in the room. Experimental work is reported for a scenario consisting of meeting-room acoustic events, either isolated or overlapped with speech. From the localization results, which are compared with those from the SRP-PHAT technique, it seems that the proposed model-based approach can be an alternative to current techniques for event-based localization.

Index Terms— Source localization, acoustic event detection, sound model, simultaneous sources, beamforming

1. INTRODUCTION

In acoustic source localization (ASL), there are some well-established methods [1]. One of them relies on calculating the direct time delay of arrival (TDOA) through cross-correlations, and combines it with information regarding the microphone position to generate a maximum likelihood (ML) based spatial estimator [2]. As the estimation of an accurate TDOA is a difficult task, the performance of this method degrades drastically either at low SNR, or in high reverberant room, or in multiple source scenarios. Other methods are based on steered beamforming (SB) [3], the most popular among them being the steered response power (SRP) [1]. Either simple delay-sum beamformers or the more robust phase transform (PHAT) filtered weights are used in SRP based methods [1], [2]. The SRP-PHAT technique is more robust at low SNRs compared to the above mentioned techniques, and it has already been established as a kind of standard in ASL [4]. Additionally, a localization technique, based on ML, using a noise model, which is closely related to SRP, has been reported in [5]. Generally, the SRP based techniques use computationally intensive grid search methods to find a global maximum. In [6], [7], the authors discussed the computational issues and proposed efficient methods so that the SRP based techniques can be implemented in real time.

To estimate the position coordinates of the acoustic sources, most widely-used ASL methods use energy-like measures extracted from the microphone signals. Conversely, in this paper, the use of the information about the content of the signals is proposed. In fact, instead of relying only on energy-like measures, the probability or similarity measure delivered by a classifier is proposed. As the classifier uses models for the different sound classes, we can refer to this approach as sound-model-based (SMB) localization. In a practical situation, the identity and the time positioning of the (possibly) simultaneous sounds may be provided by an acoustic event detection (AED) system, so the sound models are shared by both AED and ASL systems.

By discretizing the space in the room, a set of beamformers, based on a frequency invariant null-steering approach, is used to nullify, up to some extent, the signals coming from the discretized positions. Based on a set of statistical models, for each multi-channel signal we have a set of likelihood values, each one being the likelihood corresponding to a specific position in the room and a specific event class. Then, a maximum-a-posteriori (MAP) criterion is applied to estimate the optimal position of each event source in the room space. The processing scheme of this proposed ASL system is similar to the one presented in [8], [9] for acoustic event detection.

In the experimental work, we contextualize the SMB method in our smart-room, where small T-shaped microphone arrays are distributed on the walls. Experiments are carried out with a concrete meeting-room scenario with one or two simultaneous sources, and using a database collected in the smart-room. The localization results obtained with the proposed SMB method using all the six 3-microphone arrays in the room is compared to those of a baseline SRP-PHAT based localization system working in an event-based mode.

The proposed ASL system is described in Section 2. Experimental work is reported in Section 3, and a conclusion is given in Section 4.

2. SOUND-MODEL-BASED LOCALIZATION

Herewith, it is assumed that the identities and the end-points of a set of acoustic events are known. The acoustic events may occur either isolatedly or simultaneously in time.

The proposed system is shown in Fig. 1. Let us assume a room with a set of K microphone arrays which can be located arbitrarily; for deployment, this is an advantage with respect to using spatially-structured array configurations. The 2-D room space is divided into a set of P pre-defined small-area cells. Note

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

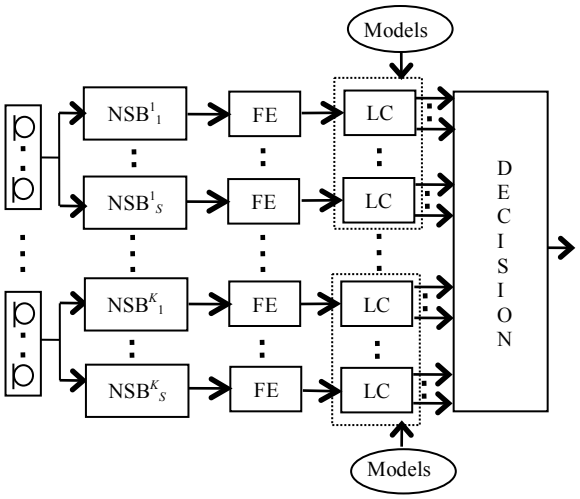

Figure 1: Proposed acoustic source localization system

that the vertical coordinate is not considered in this study, though the proposed system could easily include it. For each microphone array, there is a set of P null-steering beamformers (NSB), each one attenuating the signals from all directions except the direction corresponding to the center of one of the cells.

The output signal of each beamformer enters a classification system. After feature extraction (FE), a likelihood score (LC) is computed for each of the considered event classes (e.g. hypothesized from an external AED system), by using previously trained acoustic event models (that may be the same models used by the AED system). Finally, a decision module carries out the localization of the events by combining the likelihood scores using a MAP criterion. The proposed system is hereafter referred to as steered-beamforming sound–model-based (SBSMB) localization system. The beamformer design and the model based localization using MAP are presented in the two following subsections.

2.1. Frequency invariant null-steering beamforming

A null-steering beamformer (NSB) is capable of placing nulls at different positions in the sensor array patterns [10], [11]. Given the broadband characteristics of the audio signals, in order to determine the beamformer coefficients we use a technique called frequency invariant beamforming (FIB). The method, proposed in [12], uses a numerical approach to construct an optimal frequency invariant response for an arbitrary array configuration with a very small number of microphones, and it is capable of nulling several interfering sources simultaneously. As depicted in Fig. 2, the FIB method first decouples the spatial selectivity from the frequency selectivity by replacing the set of real sensors by a set of virtual ones, which are frequency invariant. Then, the same array coefficients can be used for all frequencies. An illustrative example is shown in Fig. 3; note how the beams for the angle of interest are rather constant along frequency.

2.2. MAP-based source localization

Let us assume a room with K microphone arrays, and a set of N possibly simultaneous) events E_i, $1 \le i \le N$, that belong to a set of C

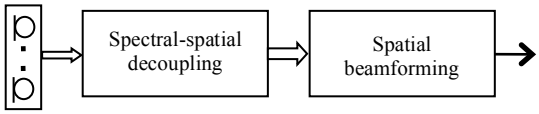

Figure 2: Frequency invariant beamforming

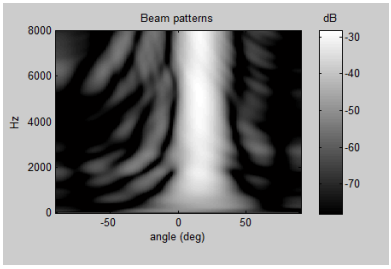

Figure 3: Example of FIB beam pattern. Angle of interest is 15 deg.

different classes. Given a grid of positions s_j, $1 \le j \le P$, in the room, for each array, there is a set of P NSBs, so that the j-th NSB is placing nulls in the directions of the P positions except that of position s_j. Therefore, there is a set of P output signals from each array processor. For a given event E_i, a set of P likelihood scores are obtained from the NSB outputs and using the model of the class E_i. The optimal position s_o^i of that i-th event out of the N events is chosen to maximize a product of posterior probabilities [13], i.e.

$$s_o^i = \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(s_j | E_i, X_k) = \underset{s_j}{\operatorname{argmax}} \prod_{k=1}^{K} p(X_k | E_i, s_j) p(s_j | E_i) / p(X_k) \quad (1)$$

3. EXPERIMENTS WITH ISOLATED AND OVERLAPPED ACOUSTIC EVENTS

In the experimental work, a meeting room scenario with a predefined set of 11 acoustic events has been considered [9], [14], [15]. Like in [9], [15], it is assumed that there may simultaneously exist 0, 1 or 2 events, and, in the last case, one of the events is always speech. In the reported experiments, we localize the isolated events (1 source) and overlapped events (2 sources). To determine the likelihoods, the acoustic events are modeled with Hidden Markov models (HMM), and the state emission probabilities are computed with continuous density Gaussian mixture models (GMM).

3.1. Meeting room acoustic scenario and database

Fig. 4 depicts our department's smart-room, with the position of its six T-shaped 4-microphone arrays on the walls. The linear arrays of 3 microphones are used in the experiments. For training, development and testing of the system, we have used, as in

[15], part of a publicly available multimodal database recorded in the smart-room. Concretely, 8 recording sessions of audio data, which contain isolated acoustic events are used. The approximate source positions of the acoustic events (AE) are shown in Fig.4. Each session was recorded with all the six T-shaped microphone arrays. The overlapped signals used for development and testing of the system were generated adding those AE signals recorded in the room with a speech signal, also recorded in the room, both from all the 24 microphones. To do that, for each AE instance, a segment with the same length was extracted from the speech signal starting from a random position, and added to the AE signal. The mean power of speech was made equivalent to the mean power of the overlapping AE. That addition of signals produces an increment of the background noise level, since it is included twice in the overlapped signals; however, going from isolated to overlapped signals the SNR reduction is slight: from 18.7dB to 17.5dB. Although in our real meeting-room scenario the speaker may be placed at any point in the room, in the experimental dataset its position is fixed at a point at the left side (SP, in Fig. 4). All signals were recorded at 44,1 kHz sampling frequency, and further converted to 16 kHz.

3.2. Event source localization

In the reported experiments, the steered-beamforming sound model based (SBSMB) system depicted in Fig. 1 is used to localize either one or two simultaneous acoustic event sources in the room environment. The nulls of the beamformers are placed in the (all but one) directions of the centers of the pre-defined cells. To facilitate real time processing, a relatively large cell: 0.6x0.8m has been considered. Though a larger cell reduces the resolution of the ASL in the room, it also reduces the number of beamformers required, which in turn ensures less computational load. In the proposed system, the beamformers are designed to work with the horizontal row of 3 microphones available in each array of the smart-room.

In the feature extraction block of the SBSMB system depicted in Fig 1, a set of audio spectro–temporal features is computed for each signal frame. Frames are 30 ms long with 20 ms shift, and a Hamming window is applied. We have used frequency-filtered log filter-bank energies (FF-LFBE) for the parametric representation of the spectral envelope of the audio signal [16]. For each frame, a short-length FIR filter with a transfer function $z\text{-}z^{-1}$ is applied to the log filter-bank energy vectors and end-points are taken into account. Here, 16 FF-LFBEs along with their 16 first temporal derivatives are used, where the latter represents the temporal evolution of the envelope. Therefore, the dimension of the feature vector is 32.

The HTK toolkit is used for developing the HMM-GMM based classifier [17]. There is one left-to-right HMM with three emitting states for each AE. 32 Gaussian components with diagonal covariance matrix are used per state. Initially, each HMM is trained, with the standard Baum-Welch algorithm, using the signals for a particular array. For each array, the likelihoods are computed by using the same set of acoustic event models for all the beamformer outputs.

Given an event class, the optimal source position is obtained by maximizing the probability resulting from product-rule combination of posteriors over all microphone-arrays, as indicated by Eq. (1). All the positions are assigned flat prior probabilities in the reported tests.

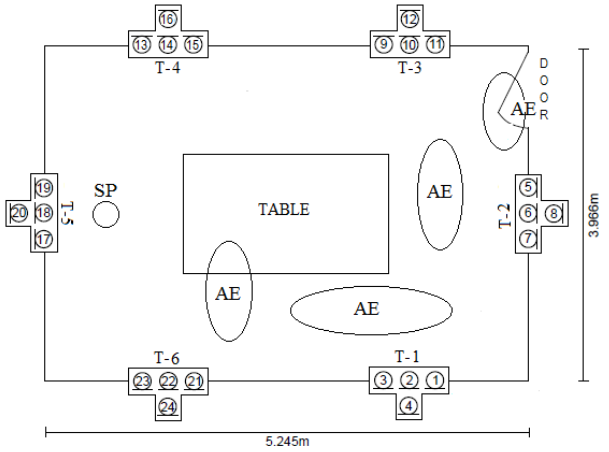

Figure 4 : Smart-room layout, with the positions of microphone arrays (T-*i*), acoustic events (AE) and speaker (SP)

3.3. Proposed metrics

To test the performance of the model based localization system, two metrics are used. 1) Acoustic source localization cell error (*Cell error*), which is defined as the quotient between the number of localization errors and the total number of event occurrences in the testing database. For an event E_i, a localization error occurs when the cell assigned to the true position is not the same as the one estimated by the ASL system. The true position for each event was obtained from visual inspection during the recording of the signal. 2) Root-mean-squared error for localization (*RMSE*):

$$RMSE = \sqrt{\frac{1}{N_e}\sum_{i=1}^{N_e}\left(\frac{\left|x_i^{test} - x_i^{ref}\right|}{\Delta x} + \frac{\left|y_i^{test} - y_i^{ref}\right|}{\Delta y}\right)^2} \qquad (2)$$

where (x_i^{test}, y_i^{test}) is the 2-D estimated position of the test event, and (x_i^{ref}, y_i^{ref}) is its corresponding reference (true) position. Δx and Δy are the separations along x and y axis of the pre-defined positions which are considered by quantizing the room space. N_e is the total number of event samples in the testing session.

3.4. Results and discussion

The testing results are obtained with all the 8 sessions (S01-S08) with a leave-one-out criterion, i.e. we recursively keep one session for testing, while all the other 7 sessions are used for training. Table 1 shows the results obtained with two metrics for the proposed SBSMB system. As a comparison, in the same table the result for a SRP-PHAT localization system has also been presented, which consists of exploring the space, searching for the maximum of the global contribution of the PHAT-weighted cross-correlations from all the microphone pairs [1], [2]. Instead of a grid-search, which requires functional evaluation on a fine grid throughout the room, a stochastic region contraction is used to find the global maximum as presented in [6], [18], to facilitate a real-time working environment. The

results with the two metrics from Sub-section 3.3, averaging over all AEs (excluding speech) in the 8 testing datasets, are obtained using all the six arrays (T1 to T6) available in the room.

The results obtained with the SBSMB system consider flat values for both $p(s_j|E_i)$ and $p(X_k)$. It is worth noticing that, in the proposed method, an event-based approach is followed, which means the localization is performed from a whole event instead of localizing in a frame-by-frame basis.

Due to that event based approach, it is assumed that during the whole event the acoustic source is not moving along space. For that reason, 'steps' and 'chair moving' events are kept out from the evaluation. In addition, instead of using the AED system output to set the AE model used by the likelihood calculators in the classifier, the ground truth has been used, so the errors from the AED system are not affecting the measure of localization performance in our tests.

In the experiments with one source (a non-speech acoustic event), the proposed system shows a slightly lower cell error rate than the conventional SRP-PHAT system. The performance scores for the acoustic events in the overlapped case (when an acoustic event is overlapped with speech), with both the SBSMB system and the SRP-PHAT system, are presented in Table 2. Both techniques have been used to localize the AE source when there are two sources present in the signals. The SRP-PHAT technique has been scored by looking at the two main peaks in the resulting acoustic map. The proposed SBSMB system clearly outperforms the SRP-PHAT based system. Notice that, for the SBSMB system, the cell error rate in the two-source case is only around 6% (relatively) higher than that of the one-source case.

Table 1: Performance comparison of the ASL systems for the isolated (one-source) case

	SBSMB	SRP-PHAT
Cell error (%)	13.4	13.8
RMSE	0.41	0.44

Table 2: Performance comparison of the ASL systems for the overlapped (two-source) case

	SBSMB	SRP-PHAT
Cell error (%)	14.5	29.1
RMSE	0.53	1.5

4. CONCLUSION

A novel approach for acoustic source localization based on models of the sounds has been presented which combines a set of beamformers and a MAP based decision. When tested in a meeting-room scenario, the one-source localization performance of the proposed system is slightly better than that of the widely used SRP-PHAT based system, while it is significantly better in the more complex two-source scenario, provided the exact information about classes and time end-points is available. Note that, unlike the SRP-PHAT system, the SBSMB system requires the identities and the time end-points of the events. However, it may take advantage of the a-priori probabilities of the predefined positions for each event class, though they were not used in the experiments. In summary, the presented SBSMB localization technique can be an alternative for localization in a multiple source scenario when it works together with an acoustic event detection system, with the additional advantage that both use the same framework. Future work will be devoted to design a combined system that uses a joint approach for localization and recognition, which does not need any assumption about identities or positions.

5. ACKNOWLEDGMENTS

This work has been supported by the Spanish project SARAI (TEC2010-21040-C02-01). Thanks are given to Carlos Segura for his helpful comments.

6. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, New York: Springer, 2001.

[2] M. Omologo and P. Svaizer, "Use of the cross-power-spectrum phase in acoustic event location," *IEEE Trans. Speech and Audio Processing.* 1993, 5, 288–292.

[3] J. Dmochowski and J. Benesty, "Steered Beamforming Approaches for Acoustic Source Localization," *Speech Processing in Modern Communication*, 1st Eds., I. Cohen, J. Benesty, and S. Gannot, Eds.; Springer-Verlag Berlin Heidelberg, vol. 3, pp. 307–337, 2010.

[4] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in low noise, reverberant environments?" *Proc. ICASSP*, Las Vegas, USA, 2008.

[5] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. on Multimedia*, vol. 10, pp. 538-548, 2008.

[6] M. F. Berger and H. F. Silverman, "Microphone array optimization by stochastic region contraction," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 39, no. 11, pp. 2377-2386, 2002.

[7] J. Dmochowski, J. Benesty, and S. Affes, "A Generalized Steered Response Power Method for Computationally Viable Source Localization," *IEEE Trans. on Audio, Speech and Language Process*ing, vol. 15, pp. 2510–2526, 2007.

[8] R. Chakraborty, C. Nadeu, and T. Butko, "Detection and positioning of overlapped sounds in a room environment", *Proc. Interspeech*, Portland, *USA,* 2012.

[9] R. Chakraborty and C. Nadeu, "Real-time multi-microphone recognition of simultaneous sounds in a room environment", *Proc. ICASSP*, Vancouver, Canada, 2013.

[10] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4-24, April, 1988.

[11] O. Hoshuyama, and A. Sugiyama, "Robust Adaptive Beamforming", in *Microphone Arrays: Signal Processing Techniques and Applications*. Ed. M. Brandstein and D. Ward. New York: Springer, 2001.

[12] L.C. Parra, "Steerable Frequency-Invariant Beamforming for Arbitrary Arrays", *Journal of the Acoustical Society of America*, 119 (6), pp. 3839-3847, June, 2006.

[13] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley-Interscience, 2004.
[14] A. Temko, C. Nadeu. D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," in *Computers in the Human Interaction Loop*, A. Waibel, R. Stiefelhagen, Eds., Springer, pp. 61-73, 2009.
[15] T. Butko, F. Gonzalez Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: online implementation in a smart-room", *Proc. EUSIPCO*, Barcelona, Spain, 2011.
[16] C. Nadeu, D. Macho, and J. Hernando, "Frequency & time filtering of filter-bank energies for robust HMM speech recognition", *Speech Communication*, vol. 34, pp. 93-114, 2001.
[17] S. Young, et al., *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.
[18] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," *Proc. ICASSP*, Hawaii, USA, 2007.