REAL-TIME MULTI-MICROPHONE RECOGNITION OF SIMULTANEOUS SOUNDS IN A ROOM ENVIRONMENT

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

Time overlapping of acoustic signals, which so often occurs in real life, is a challenge for current state-of-the-art sound recognition systems. In this work, we propose an approach for detecting, identifying and positioning a set of simultaneous acoustic events in a room environment, using multiple arbitrarily-located microphone arrays, and working in real time. Assuming a set of estimated acoustic source positions, the use of a frequency invariant null-steering beamformer for each position and each array yields a set of signals which show different balances among the various acoustic sources. For each signal, a model-based likelihood computation is carried out to obtain a matrix of likelihood scores. Then a MAP criterion is used to jointly detect the event classes and assign each of them to a given source position. Experimental results with two sources, one of which is speech, and two three-microphone linear arrays are reported, and a comparison with alternatives approaches is carried out.

Index Terms— Sound recognition, acoustic event detection, overlapped events, microphone arrays, null-steering beamforming.

1. INTRODUCTION

The detection of the presence and the identity of the diverse acoustic events that occur in a room environment may help to automatically describe the social and human activities that take place in it, and also to increase the robustness of speech processing systems. After the CLEAR'07 international evaluations, where acoustic event detection (AED) was carried out with meeting-room seminars, it became clear that time overlapping of acoustic events is a major source of detection errors [1].

The detection of overlapping events may be dealt with different approaches, either at the signal level, the feature level, the model level, etc. In [2], a model based approach was adopted for detection of events in a meeting-room scenario with two sources, one of which is always speech, and the other one is a different acoustic event from a list of 11 pre-defined events. That approach is used in the current real-time system implemented in our smart-room at the UPC, which includes both AED and acoustic source localization (ASL) [3].

However, the model based approach is hardly feasible in multi-source scenarios where either the number of events or the number of simultaneous sources is large, since all the possible combinations of events have to be modeled. In such case, the problem can be tackled at levels other than the model one. For instance, in [4], non-negative matrix factorization (NMF) is used at the front end for separating up to 4 simultaneous acoustic sources; and convolutive-NMF is used in [5] to deal with noisy acoustic events. Recognition of overlapped events at the feature level is attempted in [6-7]. In this work, we propose an alternative computationally efficient approach, which is based on signal separation by using multiple linear microphone arrays that are composed of a small number of microphones.

Assuming a set of P hypothesized source positions (e.g. provided by the ASL system), a set of P beamformers is used to separate up to some extent each hypothesized source from the others. Using those (partially) separated signals, acoustic event detection is carried out using a maximum-a-posteriori (MAP) criterion. Moreover, each hypothesized event is assigned to a given source position using the same framework. The beamformers are based on a frequency invariant null steering approach.

Experiments are carried out with the concrete meeting-room scenario mentioned above, using a database collected in the own smart-room. Results obtained with one array are compared with the ones from the model based approach, and also with those from a statistical blind source separation (BSS) technique based on the deflation method, which was already used in some initial experiments reported in [8]. Although the proposed technique shows slightly lower recognition results than the other two, it does not need the posterior assignment of event hypothesis to source positions that the other techniques require. Additionally, in the experiments it is observed how the use of an additional array further improves both the recognition accuracy and the position assignment accuracy.

The system for detection and position assignment of simultaneous acoustic events is described in Section 2. Experimental work is reported in Section 3, and a conclusion is given in Section 4.

2. RECOGNITION OF OVERLAPPED EVENTS BY MULTI-ARRAY SIGNAL SEPARATION

We start assuming that several acoustic source positions are provided. They may have been estimated by a localization system that uses the available set of microphone arrays. In our approach, the arrays can be located arbitrarily. For deployment, this is an advantage with respect to using spatially structured array configurations.

As shown in Fig. 1, in the proposed system, firstly the multi-channel signal collected by each of the microphone arrays is driven to a set of null-steering beamformers (NSB). Each beamformer is

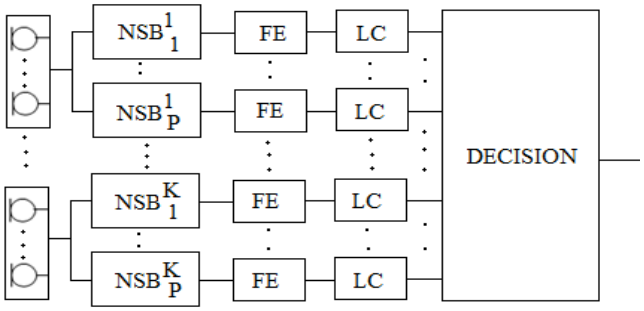

Fig. 1: Scheme of the whole detection system using K arrays.

Feature extraction (FE) is applied at the output of each beamformer, to subsequently compute a likelihood score (LC), by using previously trained models of the acoustic event classes. At last, a decision module carries out the detection of the event identities by integrating the likelihood scores using a MAP criterion. Both the beamformer design and the MAP detection are presented in the two following subsections.

2.1. Signal separation with frequency invariant null steering beamforming

Null steering beamforming (NSB) allows us to design a sensor array pattern that steers the main beam towards the desired source, and places nulls in the direction of interferent sources [9]. 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 [10], 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 interferent 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 null beams are rather constant along frequency.

Indeed, in our case we cannot expect with this approach a perfect separation of the different mixed signals at the output of the NSB, since we use a small number of microphones per array, and also because of echoes and room reverberation.

2.2. Acoustic event detection and position assignment

In our work we follow a detection approach that is based on classification. As the silence class is used, when the system is running along time and it outputs a non-silence hypothesis for a given steering direction in the array, it is decided that an event is detected at the target position in that direction. Consequently, we will deal in this section with a classification problem.

A MAP criterion is used in our system. 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).

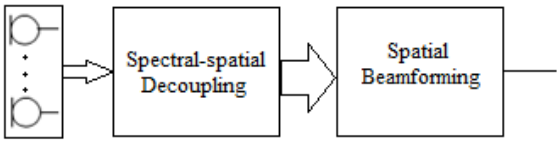

Fig. 2: Frequency invariant beamforming.

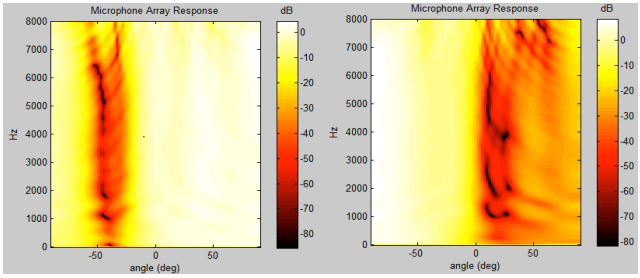

Fig. 3: FIB for two sources. The right-hand beamformer pattern shows a broader null than the left-hand one.

Let's assume we have a set of C classes, a set of P acoustic source positions, and a set of K microphone arrays. Each array steers a NSB to each of the source positions while nulling the others. So from array processing, we have a set of PK output signals, and after likelihood computations, we have a PxK-dimensional matrix of likelihood scores. We will assume also that each class c_i has a prior probability $p(c_i)$, and each estimated source position s_j has an associated probability $p(s_j)$. The latter may be provided by the ASL system.

Performing null steering beamforming with the k-th microphone array, which has at its input the multi-channel signal X_k (notice that, to simplify notation, we do not consider time indices), P output signals will be obtained, one for each NSB pattern. Let's denote with s_j the NSB that has the position s_j as target and the other positions as nulls. We want to determine the posterior probability of a given class c_i for that k-th array through all the P NSBs (note that our NSBs only separate the signals partially, so a class actually produced at position s_j may still be observed in all the NSBs that do not steer at s_j). By using the product combination rule [11] (i.e. assuming the output signals of the beamformers are independent), we have

$$p(c_i \mid X_k) = \prod_{j=1}^{P} p(c_i|s_j,X_k)p(s_j) = \prod_{j=1}^{P} p(X_k|c_i,s_j)p(c_i)p(s_j) / p(X_k) \quad (1)$$

where $p(X_k \mid c_i, s_j)$ is the likelihood of class c_i obtained from its corresponding HMM-GMM model.

For combining the posterior probabilities from the various microphone arrays, we will use here again the product combination rule, so the optimal class c_o will be obtained with

$$c_o = \underset{c_i}{\operatorname{argmax}} \prod_{k=1}^{K} p(c_i|X_k) \quad (2)$$

This beamforming based approach for AED allows to easily assign the optimal class to one of the given source positions. In fact, the optimal position will be chosen as the one steered by the beamformers whose outputs show a maximum product of posteriors over all arrays given the optimal class:

$$s_o = \underset{s_j}{\text{argmax}} \prod_{k=1}^{K} p(s_j|c_o, X_k) = \underset{s_j}{\text{argmax}} \prod_{k=1}^{K} p(X_k|c_o, s_j) p(s_j) / p(X_k) \quad (3)$$

3. EXPERIMENTS WITH A TWO-SOURCE SCENARIO

In our experimental work, we consider a meeting room scenario with a predefined set of 11 acoustic events plus speech [1-3]. Like in [3], there may exist either 0, 1 or 2 simultaneous events, and, in the last case, one of the events is always speech. However, the reported experiments correspond to the case of two overlapped events, since it is the most general one.

3.1. Acoustic scenario and database

Fig. 4 shows the UPC's smart-room, with the position of its 6 T-shaped 4-microphone arrays on the walls. We use only the linear arrays of 3 microphones in our experiments. For training, development and testing of the system, we have used, as in [3], part of a publicly available multimodal database recorded in the same smart-room. Concretely, we use 8 recording sessions of audio data which contain isolated acoustic events. The approximate source positions of the acoustic events (AE) are shown in Fig. 4. Each session was recorded with all the 6 T-shaped microphone arrays. The overlapped signals used for development and testing of the systems 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. Acoustic event recognition

As the number of sources is P=2 in our scenario, two beamformers are used: NSB1 and NSB2. This beamformers are specific for each array. In our experiments, to set the directions of arrival of the beamformers (for both the target source and the null source) we use the positions depicted in Fig. 4, so the beamformers are not adapted to each particular AE instance, as it would be done if the output of an ASL system was used. One angle corresponds to the

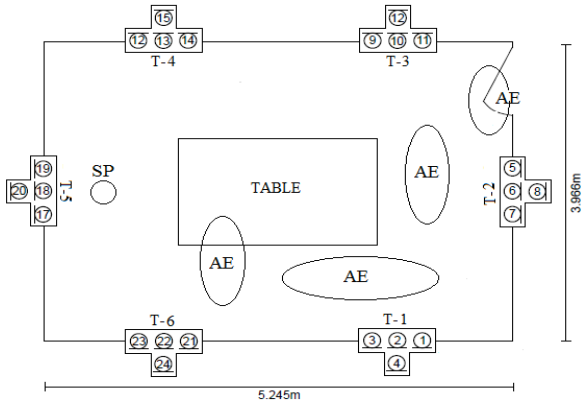

Fig. 4: Smart-room layout, with the positions of microphone arrays (T-*i*), acoustic events (AE) and speaker (SP).

speech source (the speaker's position is static), and the other one is an average of the angles from the various possible AEs positions (though some of them are changing along time, like for the 'steps' event). In this way, for a particular array, NSB1 steers to the AEs and nulls speech, and NSB2 steers to speech and nulls AEs.

In the feature extraction block of the multi-array signal separation based AED system depicted in Fig. 1, a set of audio spectro–temporal features is computed for each signal frame. The frames are 30 ms long with a 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 [12]. For each frame, a short-length FIR filter with a transfer function $z\text{-}z^{-1}$ is applied to the log filter-bank energy sequence and end-points are taken into account. Here, we have used 16 FF-LFBEs along with their 16 first temporal derivatives. Therefore, the dimension of the feature vector is 32.

The HTK toolkit is used for training and testing the HMM-GMM system [13]. There is one left-to-right HMM with three emitting states for each AE and silence. 32 Gaussian components with diagonal covariance matrix are used per state. Initially, each HMM is trained with the standard Baum-Welch algorithm using signals that have been processed with the beamformer NSB1 of a particular array. Indeed, when testing, the knowledge about the relative position (left/right) of AEs and speech is not used. For each array, the likelihoods are computed by using the same set of AE+silence models for the two beamformer outputs.

As our purpose is to compare the new signal separation based approach with other methods, in the reported experiments we actually perform classification instead of detection for all the methods, i.e. we are using the annotated time marks of the events. For the new technique, classification is carried out by combining the likelihood scores as indicated in (1). Both classes and positions are assigned flat prior probabilities in the reported tests. When using two arrays, the optimal class is obtained by integrating the posterior probabilities according to (2).

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. For the signal separation based techniques the signals for training the models are taken after the separation is done; for the model based one, overlapped signals are used.

Table 1 shows the recognition results obtained with the proposed system, averaging over all the 8 testing datasets, for two

different arrays (T4 and T6), and their combination using a product rule. It can be observed that a better result is obtained from array T4 than from array T6. And the system that combines the two arrays produces a higher accuracy.

Like in our preliminary work in [8], comparison results are obtained with two other techniques, but here using a leave-one-out strategy to increase their reliability. First, a blind source separation (BSS) technique, which is based on the deflation method [14]. In coherence with how we train the models for our proposed technique, separated signals have been used for training the corresponding BSS-based system. Second, a model based technique that uses a set of 11 models (plus silence) for the acoustic events overlapped with speech. The BSS-based system also uses three microphones (T6), and the model based system uses signal from only one microphone from the same array.

Results are shown in Table 2. The technique presented in this paper is called NSB-FIB in that table. The model based system shows a higher accuracy than the two source separation techniques. Interestingly enough, the NSB-FIB system works slightly better than the more computationally demanding BSS based system when both arrays are used. Note also that, among the three evaluated techniques, only the proposed NSB-FIB system is able to assign the hypothesized classes to the given source positions without requiring an extra system for that. Its position assignment capability is evaluated in the next sub-section.

Table 1: Recognition accuracies obtained with the proposed system

	T4	T6	T4+T6
Accuracy (%)	79.18	77.84	81.83

Table 2: Recognition accuracies of the three compared systems

	Source separation based		Model based
	NSB-FIB	BSS	
Accuracy (%)	81.83	80.75	83.6

3.3. Position assignment of the recognized events

To have a complete description of the acoustic scene in our room, there is still the need of assigning each one of the two positions to each one of the two detected events. The position assignment is done at the decision block according to (3), after the optimal class is chosen through event detection. In our particular scenario, the optimal event class (1 of 11 AEs) is assigned to one of the source positions, and the other position is assigned to speech.

A position assignment rate (PAR) metric is defined for a given AE class as the quotient between the number of correct decisions and the total number of occurrences of that AE class in the testing database [8]. The results with that metric, averaging over all AEs in the 8 testing datasets, are presented in the first row of Table 3 for two different arrays (T4 and T6) and their combination using the product rule. Again, the combination of arrays produces the best result.

We have also carried out tests by using additional models: 1) for each beamformer output, we have included a second LC module that uses a set of models trained from signals that have been processed with the beamformer NSB2 (instead of NSB1); and 2) in the decision block, instead of maximizing the product of likelihoods, we maximize the product of the likelihood ratios corresponding to each beamformer, like in [15]. As presented in the second row of the Table 3, this system modification produces a slight improvement in terms of PAR for the case of a single array, and a noticeable improvement for the combination of the two arrays as reported. We also tried that modification for the classification task; however, the accuracy changed only very slightly.

Table 3**:** Position assignment results

		T4	T6	T4+T6
PAR (%)	NSB1 trained models	89.5	89.1	90.3
	NSB1+NSB2 trained models	90.2	89.4	91.7

4. CONCLUSION AND FUTURE WORK

A new approach for computationally effective detection and positioning of acoustic events that results from the combination of a beamforming-based partial signal separation and a MAP-based decision has been presented, and has been tested in a limited scenario with two sources. Although the recognition rate of the proposed technique with one array is lower than those from the other two tested techniques, when two arrays are employed, it is higher than the recognition rate from the much more computationally demanding BSS technique. On the other hand, the model-based approach has the drawback of a limited scalability regarding both the number of event classes and the number of simultaneous sources. Furthermore, the proposed technique does not need the posterior assignment of event hypothesis to source positions that the other two techniques require.

Future work will be addressed to employ the full set of linear arrays that exist in the smart-room, and also to extend the scenario to three or more simultaneous sound sources.

5. ACKNOWLEDGMENTS

This work has been supported by the Spanish project SARAI (TEC2010-21040-C02-01). Thanks are given to Taras Butko for his help at the beginning of this work.

6. REFERENCES

[1] 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, 2009, pp. 61-73.

[2] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments", *Pattern Recognition Letters*, vol. 30/14, pp. 1281-1288, Elsevier, 2009.

[3] 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.

[4] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, “Sound event detection in multisource environment using source separation”; CHIME workshop, satellite event of *Interspeech,* Florence, Italy, 2011.

[5] C. V. Cotton and D. P. W. Ellis, “Spectral vs. spectro-temporal features for acoustic event detection”, *IEEE workshop on Application of Signal Processing to Audio and Acoustics,* New York, USA, 2011.

[6] J. Dennis, H. D. Tran, and E. S. Chng, “Overlapping sound event recognition using local spectrogram features with the generalized Hough transform”, *Proc. Interspeech*, Portland, USA, 2012

[7] J. Dennis, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification in mismatched conditions”, *Signal Processing Letters*, IEEE, vol. 18, no. 2, pp. 130-133, 2011.

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

[9] 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.

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

[11] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley-Interscience, 2004.

[12] 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.

[13] S. Young, et al., *The HTK Book (for HTK Version 3.2)*, Cambridge University, 2002.

[14] C. Simon, P. Loubaton, and C. Jutten, “Separation of a class of convolutive mixtures: a contrast function approach”, *Signal Processing*, Volume 81, Issue 4, pp. 883-887, Elsevier, 2001.

[15] R. Chakraborty, C. Nadeu, and T. Butko, “Binary position assignment of two known simultaneous acoustic sources”, *Proc. iberSPEECH2012*, Madrid, Spain, 2012.