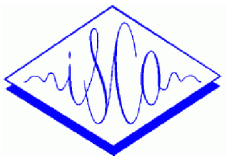
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

DETECTION AND POSITIONING OF OVERLAPPED SOUNDS IN A ROOM ENVIRONMENT

Rupayan Chakraborty, Climent Nadeu, and Taras Butko

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

10.21437/Interspeech.2012-426

Abstract

The description of the acoustic activity in a room environment has to face the problem of overlapped sounds, i.e. those which occur simultaneously. That problem can be tackled by carrying out some kind of source signal separation followed by the detection and recognition of the identity of each of the overlapped sounds. An alternative approach relies on modeling all possible overlapping combinations of acoustic events. For a spatial scene description, there is still the need of assigning each of the detected acoustic events to one of the estimated source positions. Both detection approaches are tested in our work for the case of two simultaneous sources, one of which is speech, and an array of three microphones. Blind source separation based on the deflation method and null steering beamforming are used for signal separation. Also a position assignment system is developed and tested in the same experimental scenario. It is based on the above mentioned beamformer and takes the decision based on a likelihood ratio. Both signal-level fusion and likelihood fusion are tried to combine the information from the two pairs of microphones. The reported experimental results illustrate the possibilities of the various implemented techniques.

Index Terms: Acoustic event detection, source separation, null steering beamforming, source position assignment.

1. Introduction

Systems for acoustic event detection (AED) and acoustic source localization (ASL) help to automatically describe the social and human activities that take place in a room environment, and may also increase the robustness of speech processing systems [1] [2]. After the CLEAR'07 international evaluations, where AED was carried out with meeting-room seminars, it became clear that time overlapping of acoustic events (AE) is a major source of detection errors [1]. The detection problem due to this overlapping may be dealt with different approaches, either at the signal level, at the model level, or at the decision level. In [1], the model based approach was taken for AED in a scenario with two sources, one of which is always speech. Hence, additional acoustic models were considered for each AE overlapped with speech, so the number of models was doubled. That approach is used in the current real-time system implemented in our smart-room, which includes both AED and ASL [3]. However, this approach may become unfeasible when the number of classes and the number of simultaneous sources are large. Alternatively, we can tackle the problem at the signal level by separating the signals first and then applying the AED system for isolated events.

In this work, we want to compare both alternative approaches: model-based and signal-based. Regarding the latter, we have chosen two very different source separation techniques: 1) a blind source separation technique which employs a contrast function based deflation method [4], and 2) a computationally simpler array processing based separation technique which employs null steering beamforming [7]. The advantage of these methods over the model based one is that they do not require the extra models for the overlapped signals. But, on the other hand, an additional separation block is needed at the first stage of the system. Concerning the two separation techniques, the BSS technique is much more time consuming than the other, so it is not well suited to be implemented in real-time environments.

The other aspect of the work presented in this paper is position assignment (PA) of the detected events. In fact, although the estimated position of the sources is provided by an ASL system, there is an ambiguity regarding the correspondence between identified events and located sources. Our PA system consists of the beamforming-based signal separation technique mentioned above, which is applied to two pairs of microphones, followed by a likelihood-ratio based binary classifier.

2. Acoustic scenario and database

Figure 1 shows the smart-room we have at the Universitat Politècnica de Catalunya (UPC), with the position of its 6 T-shaped 4-microphone arrays on the walls. The total number of considered acoustic event classes is 12, including speech [3]. In the working scenario, it is assumed that speech is always produced at one side of the room (left or right), and the other AE are produced at the other side. Note that the left upper corner is taken as the reference origin.

For the offline design of the system, whose real-time version is posteriorly implemented in the room, a database is needed. We recorded it using the spatial distribution of AE sources, including the speech source, depicted in Figure 1. The position of the speech source was rather fixed, but the other AE were produced within broad areas of the room layout. Note that, though in our real meeting-room scenario the speaker may be placed at either left or right side of the room, in the database its position is fixed. This will not constraint the usefulness of the results, because the system will not make use of that knowledge. As in [3], we have used for training, development and testing up to 8 sessions of audio data with isolated acoustic events. Each session was recorded with all the 6 T-shaped microphone arrays (24 microphones). The overlapped signals of the database were generated adding those AE signals recorded in the room with speech signal, also recorded in the room from all the 24 microphones. To do that,

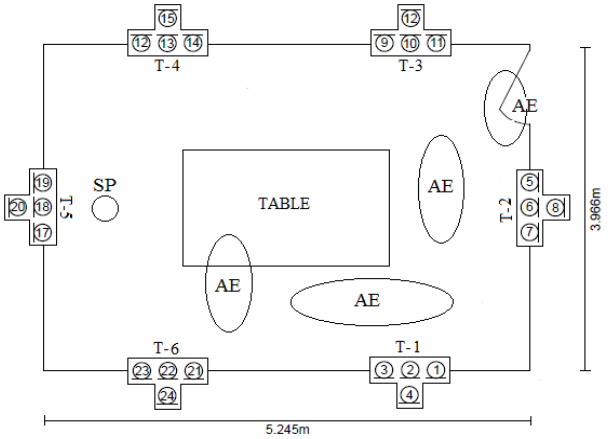


Figure 1: AE source positions and speaker (SP) position in the UPC's smart-room

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.

3. Detection of overlapped acoustic events

In a meeting room scenario it is highly probable that two or more acoustic events occur simultaneously. This concurrent occurrence makes the detection of each event even more difficult. Two approaches are described in the following subsections to deal with this problem. All the techniques will be presented for the general case, though they will be applied in this paper to the concrete scenario described in section 2.

3.1. Model based AED

In the model based approach, the AED system requires a model for each class whether it is an isolated or an overlapped event. So there must be a different model for each possible combination of classes. This approach does not require a prior separation of the two overlapped signals, but requires a number of models that may be too large. In our particular meeting-room scenario, however, the approach is feasible because we consider 11 AE that may be overlapped only with one class, speech [1][3].

3.2. Source separation based AED

The problem of overlapping can be solved at the signal level by source signal separation, i.e. the signals are first separated and then detected. Two very different techniques are considered in this work: 1) a blind source separation (BSS) technique which employs a contrast function based deflation method [4], and 2) a computationally simpler array processing based separation technique which employs null steering beamforming. The first one is expected to produce better separation than the second. The latter produces a partial separation, but its relatively low computational load may be very useful in the context of online or real-time implementation and for a specific application.

3.2.1. Source separation based on deflation

In blind source separation, signals are separated from the set of mixed signals with or without the aid of information about the source signal or the mixing process. For our work, we have selected an iterative BSS technique where the source signals are extracted from the mixtures one by one [4] [9]. The main assumptions are: the signals are stationary and statistically mutually independent, there are more sensors than sources, and the mixing system is a FIR filter. After separation, the output signals correspond in any order to the source signals passed through a scalar filter. If the sources are temporally independent and identically distributed, the scalar filter further reduces to a delay and scaling factor. Here we will use a deflation based BSS approach which consists of using a contrast function to transform the original problem into an optimization problem [10]. There are several contrast functions which can be used for this optimization. In this work, to reduce time complexity, we have used a quadratic contrast function with 4^{th} order cumulants, like in [10].

3.2.2. Source separation based on beamforming

In this second approach, source separation is based on signal processing using a null steering beamformer (NSB) [7]. In the first stage, the NSB adapts the microphone array pattern by steering the main beam towards the desired source and placing nulls in the directions of the interference sources. Thus the contribution of one of the simultaneous sounds to the beamformer output is expected to be lower than its contribution to the beamformer input. In the case of two sources, we will have two NSB in parallel, so each of the two outputs will nullify a different source signal. Indeed, beamforming is based on the prior knowledge of the direction of the desired and interference sources, which can be provided by an ASL system. Thus, each NSB has two inputs: 1) the multimicrophone signal and 2) position coordinates or direction of arrival (DOA) of the sources. In the reported experiments, a linear array of only two microphones is used to design a first-order NSB.

3.3. Experiments

In the first stage of the AED system based on any one of the previously mentioned techniques, the feature extraction block extracts a set of audio spectro–temporal features for each signal frame. In the experiments, the frame length is 30 ms 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. 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, we have used 16 FF-LFBEs along with their 16 first temporal derivatives, where the latter represents the temporal evolution of the envelope. Therefore, the dimension of the feature vector is 32.

In the recognition stage, we have used a hidden Markov model (HMM) based classifier, where Gaussian mixture models (GMM) are used to compute state emission probabilities [5]. The HTK toolkit is used for training and testing our HMM-GMM system [6]. There is one left-to-right HMM with three emitting states for each AE. This HMM topology showed the best results with cross validation on the development data. The observation distributions of these states are Gaussian mixtures with continuous density. 64 Gaussian components with diagonal covariance matrix are used per model. Each HMM is trained with the signal segments belonging to the corresponding event class using the standard

Baum-Welch training algorithm [5]. A total of 22 HMMs are trained, one for each isolated AE class and one for each AE class overlapped with speech. The Viterbi algorithm is used for testing.

To test the performance of the AED systems, we have used the same metric (AED-ACC) used in [1] [3], and also the same partition for training and testing datasets used in [3]. 7 recording sessions (S02-S08) are used for training, and the remaining session S01 for testing. For training and testing the model based AED system, both isolated and overlapped signals are used. For training the two source separation based techniques, separated signals are used, to avoid a mismatch between training and testing. However, as it has been observed that the inclusion of overlapped signals during training increases the performance accuracy of the source separation systems, overlapped signals are also included in training, like was done for the model based AED system.

The detection accuracy of the three systems is presented in Table 1. It can be seen that the model based AED system shows the best result. Interestingly enough, in our application, the NSB based system performs only slightly worse than the much more complex BSS based system.

Table 1: Detection accuracy of the various AED systems

	Model based	BSS based	NSB based
Average AED rate (%)	89	82	81

4. Binary position assignment

After applying AED, the identity of the detected AE (either one or two) is known. On the other hand, the ASL system provides either one or two source positions. Thus, in the most general case, we have two detected events, i.e. *E* (one of the 11 possible AE) and "sp", and two source positions: P_1 and P_2. 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 AE. In this section we want to design a system that can be deployed in real time in the room to resolve that ambiguity in the correspondence between AE labels and acoustic source positions.

The whole set of systems is depicted in Figure 2. The two-source AED system and the two-source ASL system are those described in [3]. The first one uses the model based approach (Section 3.1), which showed the best result in our task. And the second, which employs all 24 microphones, is based on the SRP-PHAT localization method. In this section we will assume there are always two simultaneous events, so at the output of the AED system we need only the hypothesized identity of the non-speech AE, as indicated in Figure 2 by *E*. The ASL system provides an estimate of the two source positions. The position assignment (PA) block actually is a binary classifier that assigns *E* to either P_1 or P_2.

The PA system, which is shown in Figure 3, has at its front-end two NSB, which work in parallel, like in the AED system presented in Section 3.2.2. Each of the beamformers is followed by a likelihood computation, which uses the HMM model corresponding to the acoustic event *E*. Finally, a decision block makes the assignment based on the two computed likelihoods.

In our work, beamforming is based on the knowledge of the directions of the desired and the interference source. A 1st

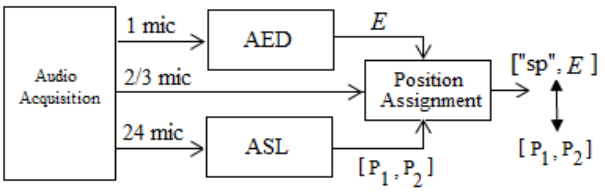


Figure 2: Block diagram of the whole system

order beamformer is designed from two microphones [8]. When using the 3-microphone array, we have considered two pairs of microphone signals. The outputs of the two 1st-order beamformers are either combined at the signal level or at the classifier level. The latter is the option taken in Figure 3. The weight vectors $w_{i,j}$ in each of the NSB are calculated from the positions provided by the ASL block, where *i* indicates the number of the beamformer, and *j* is the number of the pair of microphone for each beamformer.

The classifier consists of a feature extraction block at the first stage and a likelihood calculator at the second stage, like in the AED system presented in Section 3.2.2. After combining the two likelihoods with the product rule at each of the two parallel paths, the outputs are fed to the decision block which, by comparing the combined log-likelihood values, assigns one of the two positions provided by the ASL system to the detected AE and the other position to speech. Hence, it takes a binary decision based on a log-likelihood ratio. The path with the largest log-likelihood output is taken as the one where a null is placed at the speech source direction, so the speech source position is decided, and the other position corresponds to the AE source.

4.1. Design of the PA system and metrics

In our experimental study, we have used one beamformer, let's say NSB1, for nulling speech, and the other beamformer, let's say NSB2, for nulling the acoustic event. As the classifier is trained with isolated acoustic events, it is expected to get comparatively higher log-likelihood from the output of NSB1 than from the output of NSB2. In case a decision is made that speech corresponds to the right side source position, and the AE to the left side one, it is counted as a correct decision (since the speech source position always is, in the database, at same right hand position, as in Figure 1). The opposite one is counted as an error.

To design and evaluate the performance of the system, we define the Position Assignment Rate (PAR) metric 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. We have also considered a second metric called Diff_LL, which is defined as the difference between the two log-likelihood values computed by the calculators in the case of having a correct assignment. While maximization of the PAR is our main criterion for evaluation, Diff_LL has also been used as a secondary indicator of quality. In fact, as the beamformers assume a single frequency signal, when tuning that frequency *f*, sometimes occurs that the PAR for an AE is the same for several frequency values, since the number of AE occurrences is not high enough. Then, the frequency *f* that maximizes Diff_LL is chosen. In fact, that difference can be considered as an estimate of the degree of source separation carried out by the beamforming system for correct assignment decision.

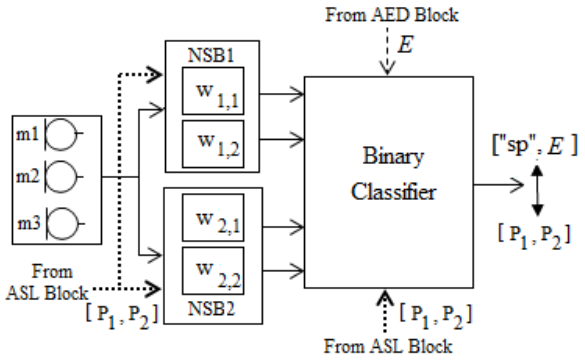


Figure 3: Position assignment system

Indeed, the beam patterns of the NSB are strongly dependent on both the DOA and the frequency *f*. We have used the one-source ASL system outputs for AE's positions; speech source position is taken from the prior knowledge gained during recording. So a different DOA is considered for each AE occurrence. Regarding *f*, as the overlapped signals are wideband, two types of implementation for the NSB are considered: 1) in the time domain, by tuning *f*, and 2) in the frequency domain. In the former approach, as we are looking to design our system for different types of acoustic events with diverse spectra, it is reasonable to choose *f* for each specific event. Hence, in the first type of implementation, the two beamformers are tuned with respect to frequency as well as DOA to get optimal results. We adopted an exhaustive search technique for empirical frequency optimization, varying the frequency from 100 Hz to 8 KHz in intervals of 100 Hz, and observing the performance of the system for each acoustic event separately. We have used data from seven sessions (S02-S08) for that frequency optimization (the remaining session (S01) is left for testing the whole system). Cross-validation is used to have a better statistical meaning. Six sessions are used for training the AE models, and the remaining session is used for testing each frequency value. The frequency that shows the best average PAR is chosen.

In this work, we have also implemented the beamforming by converting the time domain signal to the frequency domain with the DFT, and using a beamformer for each frequency bin. Moreover, as the separation between the microphones (20 cm) is well suited for an operating frequency smaller than 1 KHz [11], we have used a low pass filter at the first stage of the system to filter the input signal with 1 KHz cut-off frequency. The main advantage of the frequency domain approach is that it does not require any frequency tuning.

4.2. Experimental results

We have done the experiments with the array T-6, using first two microphones and then extending it to three. The combination of the two beamformers at the signal level and at the classifier level has been tested. For comparing these two alternatives, we used beamformers implemented in the time domain. In the experiments reported in this paper related to the evaluation of the performance of the PA system, the ground truth was used, i.e. the errors from the AED system are not affecting the measure of position assignment performance.

The experimental result for the testing session S01, while the system is trained with S02-S08 as in [3], are presented in Table 2. Both metrics are used, averaging over all acoustic event classes. Notice that the PAR score of the frequency domain system is lower than that of the time domain system. However, its performance is much higher in terms of Diff_LL, which indicates that it achieves a better NSB based signal separation when the position assignment decision is correct. On the other hand, the classifier level likelihood combination proves to be preferable to the signal level one.

5. Conclusions

In our experimental scenario, the model based approach has shown a higher AED performance than the signal separation approach. However, according to our simulation results, a simple source separation technique based on NSB is able to obtain similar AED results than a much more demanding statistical blind source separation system, assuming the source positions provided by the ASL system are accurate. Additionally, the same NSB based signal separation block can be used as a pre-processor for the system that resolves the ambiguity in assigning the source positions to the detected overlapped events. In that PA system, the combination of the two microphone-pair paths at the decision level performs better than the alternative signal level combination.

Table 2: Experimental results for the PA system

	2mic (time)	2 mic (freq.)	3 mic (sig. comb.)	3 mic (LL comb.)
PAR	90	83	91	94
Diff_LL	1.51	3.84	2.64	2.61

6. References

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

[2] X. Zhuang, X. Zhou, M. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection", *Pattern Recognition Letters*, vol. 31/12, pp. 1543-1551, Elsevier, 2010.

[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] 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.

[5] L. Rabiner, and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

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

[7] 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.

[8] H. Teutsch, and G. W. Elko, "First and second order adaptive differential microphone arrays", Proc. *IWAENC*, 2001.

[9] M. Castella, S. Rhioui, E. Moreau, and J.-C. Pesquet, "Quadratic higher-Order criteria for iterative blind separation of a MIMO convolutive mixture of sources", *IEEE Trans. Signal Processing*, Vol. 55, Issue 1, pp. 218-232, January, 2007.

[10] Marc Castella, and Eric Moreau, "A new optimization method for reference-based quadratic contrast functions in a deflation scenario", Proc. *ICASSP*, pp. 3161-3164, Taiwan, R.O.C., 2009.

[11] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Experimental evaluation of a nested microphone array with adaptive noise cancellers", *IEEE Transactions on Instrumentation and Measurement*, Vol. 53, Issue 3, pp. 777-786, June 2004.