EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Source ambiguity resolution of overlapped sounds in a multi-microphone room environment

Rupayan Chakraborty*, Climent Nadeu and Taras Butko

Abstract

When several acoustic sources are simultaneously active in a meeting room scenario, and both the position of the sources and the identity of the time-overlapped sound classes have been estimated, the problem of assigning each source position to one of the sound classes still remains. This problem is found in the real-time system implemented in our smart-room, where it is assumed that up to two acoustic events may overlap in time and the source positions are relatively well separated in space. The position assignment system proposed in this work is based on fusion of model-based log-likelihood ratios obtained after carrying out several different partial source separations in parallel. To perform the separation, frequency-invariant null-steering beamformers, which can work with a small number of microphones, are used. The experimental results using all the six microphone arrays deployed in the room show a high assignment rate in our particular scenario.

Keywords: Source position assignment; Null steering beamforming; Acoustic event detection; Simultaneous sounds; Fuzzy integral-based fusion

1. Introduction

Sound is a rich source of information. For that reason, machine audition [1] plays an important role in many applications. In particular, for meeting room scenarios, knowledge of the identity of possibly simultaneous sounds that take place in a room at a given time and their position in space is relevant to automatically describe social and human activities [2-5], to increase the robustness of speech processing systems operating in the room [6], to assist video conferencing [7], etc.

Acoustic event detection (AED) systems try to determine the identity of an occurring sound and the time interval when it is produced [2-4]. Acoustic source localization (ASL) systems estimate its position in space [5,8-11]. Both tasks become much more challenging when there exists sound simultaneity, i.e., several sounds overlapping in time and in a given room. For example, after the CLEAR'07 international evaluations [12], where AED was carried out with meeting room seminars, it became clear that time overlapping of acoustic events (AEs) was a major source of detection errors [13].

In the concrete scenario used for the experiments, a typical meeting room acoustic scene is considered, where only a person is speaking at a given time and other non-speech sounds may happen simultaneously with the speaker's voice. Therefore, we have to deal with the problem of detecting and localizing an acoustic event that may be temporally overlapped with speech. The detection of overlapping events may be tackled with different approaches, either at the signal level, at the model level, or at the decision level. In [13-15], a model-based approach was adopted for detection of events in that meeting room scenario with two sources, one of which is always speech and the other one is an acoustic event from a list of 11 predefined events. Thus, besides the mono-event acoustic models, additional acoustic models were considered for each AE overlapped with speech, so the number of models was doubled (22 in that case). That approach is used in the current real-time system implemented in the smart-room of Universitat Politècnica de Catalunya (UPC), which includes both AED and ASL [14].

In that model-based approach, a permutation problem exists. In fact, the AED system gives the hypothesized identities of the overlapped sounds, but does not associate each of them to one of the available source positions

* Correspondence: rupayan.chakraborty@upc.edu
TALP Research Center, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC), Barcelona 08034, Spain

that are provided by the ASL system. The same problem may be encountered by using other AED approaches, for instance, if a blind source separation technique is used prior to the detection of each of the separated events. To solve that source ambiguity problem, a position assignment (PA) system that performs a one-to-one correspondence between the set of source positions and the set of class labels is presented in this paper.

The proposed PA system, a preliminary version of which was presented in [15,16], consists of three stages: beamforming-based signal separation, model-based likelihood calculations, and fusion of log-likelihood ratios over the set of beamformers. Frequency-invariant null-steering beamformers are designed for the small microphone arrays existing in the smart-room. In addition, the likelihoods coming from both a speech model and an acoustic event model are combined in order to improve the system accuracy. The work presented in this paper is an extension and improvement of the work reported in [15,16]. On the contrary, to that previous published work, in the PA system reported here, all the six microphone arrays available in the room are employed in the experiments. For taking the decision, the scores obtained from each array are combined using either a product of likelihood ratios or a fuzzy integral-based fusion technique [17-19]. Experiments are carried out for the scenario described above and with a set of signals collected in the smart-room. The observed PA accuracy is larger than 95%.

The acoustic scenario is presented in Section 2 together with the signal database used for experimentation. The position assignment system is described in Section 3. Experiments are reported in Section 4, along with some practical issues about the system implementation and the used metrics. Conclusions are presented in Section 5.

2. Acoustic scenario and database

Figure 1 shows the smart-room of the UPC, with the position of its six T-shaped four-microphone arrays on the walls. The linear arrays of three microphones are used in the experiments. The total number of considered acoustic event classes is 12, including speech, as shown in Table 1. In the working scenario, it is assumed that speech is always produced at one side of the room (either left or right), and the other AEs are produced at the other side.

For the offline design of the system, whose real-time version is later implemented in the room, a database is needed. It was recorded using a spatial distribution of the AE sources and the speech source, as depicted in Figure 1. The position of the speaker was rather stable, but the other AEs were produced within broad areas of the room layout. There are eight recorded sessions (S01 to S08) of isolated AEs, where six different persons had participated and performed each AE several times. Note that, though in a real meeting room scenario the speaker may be placed at either the left or the right side of the room, in the database its position is fixed. This will not constrain the usefulness of the results, because the system will not make use of that knowledge.

As in [14], we have used for training, development, and testing up to eight sessions of audio data with isolated acoustic events. Each session was recorded with all the six T-shaped microphone arrays (24 microphones). The overlapped signals of the database were generated adding those AE signals recorded in the room with a speech signal, also recorded in the room for a single speaker 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 18.7 to 17.5 dB. The average duration of the events is 500 ms, and the reverberation time of the room is around 450 ms. Signals were recorded at 44.1 kHz sampling frequency and further converted to 16 kHz.

3. Source position assignment

The block diagram of the whole system that performs position assignment from the outputs of the acoustic event detection and localization systems is depicted in Figure 2. The model-based AED system outputs either one or two AE hypothesis. On the other hand, in the online implementation at the UPC's smart-room, the ASL system provides either one or two source positions. Hence, there are four different possibilities for mapping the one/two detected events into the one/two detected positions. As can easily be noticed, there exists an ambiguity in three out of those four possibilities. This work is focused on the most general case, where two events are detected, i.e., E (one of the 11 possible AEs) and 'sp,' and also two source positions: P_1 and P_2. Hence, the position assignment (PA) block actually is a binary classifier that assigns E to either P_1 or P_2.

If the problem of assigning the two events to the two positions is solved, the other two cases with ambiguity can also be solved using the same approach. In this section, the aim is to design a system that can be deployed in real time in the room to resolve that ambiguity in the correspondence between detected AEs and acoustic source positions.

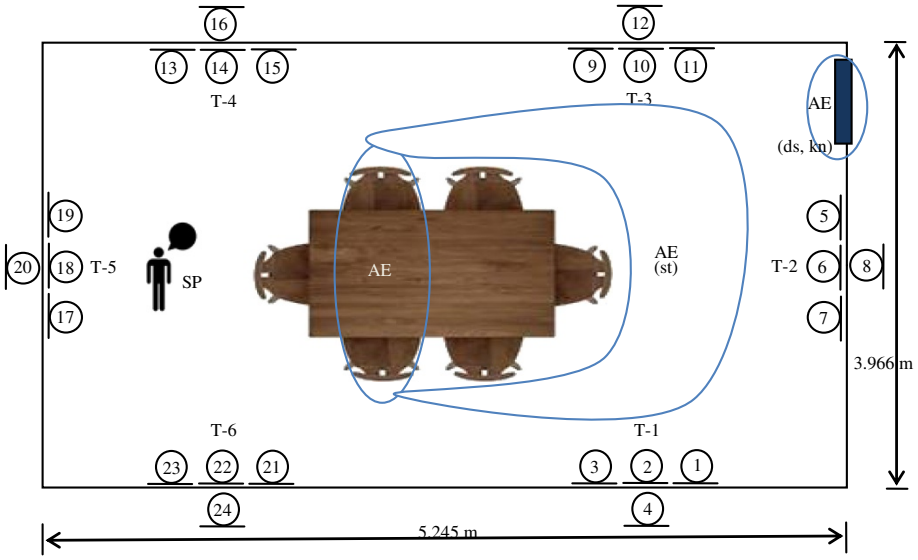

Figure 1 Smart-room (4 m × 5.3 m × 4 m) layout, with the positions of microphone arrays (T-*i*), acoustic events (AE), and speaker (SP).

3.1. The AED and ASL subsystems

The two-source AED system included in Figure 2, which was developed in previous work [13], employs a model-based approach with one microphone. All accepted sound combinations are modeled, i.e., the AED system has a model for each class whether it is an isolated event or a combination of events. 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 11 AEs are considered, which may be overlapped only with one class, speech, so only 22 models are required [14,15]. The ASL system, also developed in previous works [14], is based on the steered response power with phase transform (SRP-PHAT) algorithm, which uses 24 microphones available in the room.

Table 1 Acoustic classes and their number of occurrences

Label	Event type	Number of occurrences
[ap]	Applause	88
[cl]	Spoon/cup jingle	96
[cm]	Chair moving	242
[co]	Cough	90
[ds]	Door open/slam	256
[kj]	Key jingle	82
[kn]	Door knock	79
[kt]	Keyboard typing	89
[pr]	Phone ring	101
[pw]	Paper work	91
[st]	Step	205

3.2. The position assignment system

The scheme of the PA system, which is shown in Figure 3 for one array, has at its front-end two null-steering beamformers (NSBs), which work in parallel. The main beam of each NSB is steered towards the desired source and a null is placed in the direction of the interfering source, so each NSB will nullify a different source signal. 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. Indeed, beamforming is based on the prior knowledge of the direction of the desired source and the interferent source, which can be provided by an ASL system. Thus, each NSB requires two inputs: (1) the multi-microphone signal and (2) the position coordinates or direction of arrival (DOA) of the sources.

Each of the beamformers is followed by feature extraction (FE) and likelihood computation (LC). In this work, hidden Markov model and Gaussian mixture model (HMM-GMM) are employed, for both acoustic events and speech. Given the AE class E, the model for E and the model for speech (sp) are needed for the likelihood computations. Finally, a decision block makes the assignment based on the computed log-likelihoods.

3.3. Null-steering beamforming

Null-steering beamforming is one of the earliest, but potentially very useful beamforming techniques. It belongs to a class of very popular and widely used beamforming techniques called multiple side lobe cancellers (MSC) [20-22]. NSB adapts the sensor array pattern by steering the main beam towards the desired source and placing nulls in the direction of the interfering sources [23]. The

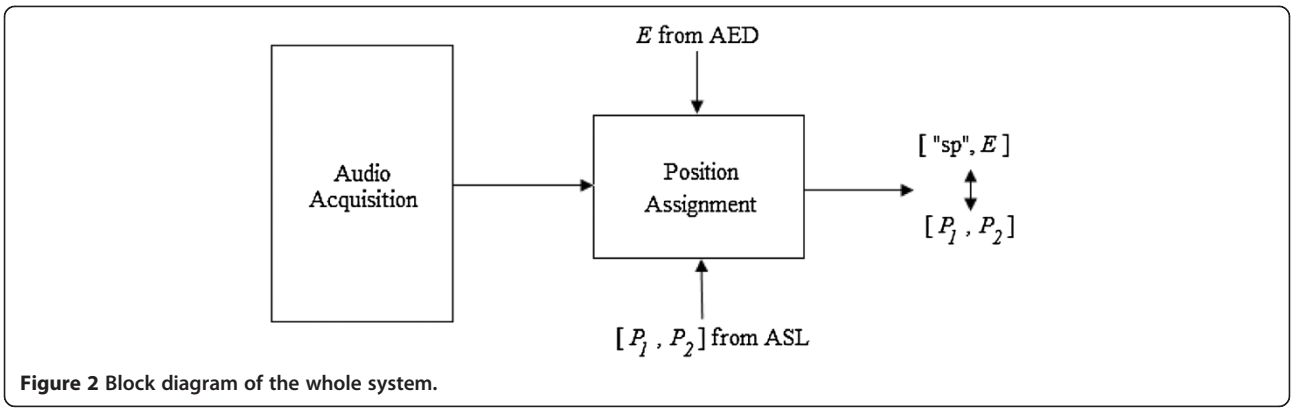

Figure 2 Block diagram of the whole system.

solution for the weight matrix in this type of beamformer is achieved by setting to unity the desired response at the direction of the target sound and setting it to zeros at the direction of the interfering sources. In our particular scenario, only one interfering source has to be nulled, so only two microphone signals are required to get a solution for the weight matrix. However, as in each T-shaped array of the room, there are three linearly spaced microphones; all three of them will be used in the experiments.

Given the broadband characteristics of the audio signals, to determine the beamformer coefficients, a technique called frequency-invariant beamforming (FIB) is used [24-27]. The method, proposed in [25] and [26], 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 Figure 4, 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 alternative frequency-dependent approach was explored for a single array in our previous works [15,16], using either a time domain implementation or a frequency domain implementation. However, in spite of carrying out a careful frequency tuning, the obtained PA accuracy was just slightly higher than the one from the FIB-based system for the time domain implementation [16]. On the other hand, the alternative FIB technique does not require frequency tuning, and thus, it is less dependent on the concrete scenario. For those reasons, FIB was chosen in the work reported here.

3.4. Single-array classification stage

As shown in Figure 3, the classification stage of the PA system with a single array consists of feature extraction, followed by log-likelihood calculation and a binary decision block. Features are extracted from the audio signals with a frame length of 30 ms and a frame shift of 20 ms. As features, frequency-filtered log filter-bank energies (FF-LFBEs), which were developed for speech recognition, are used [28]. These features are uncorrelated, similarly to the conventional MFCC features, but unlike the latter, the FF-LFBE features keep the frequency

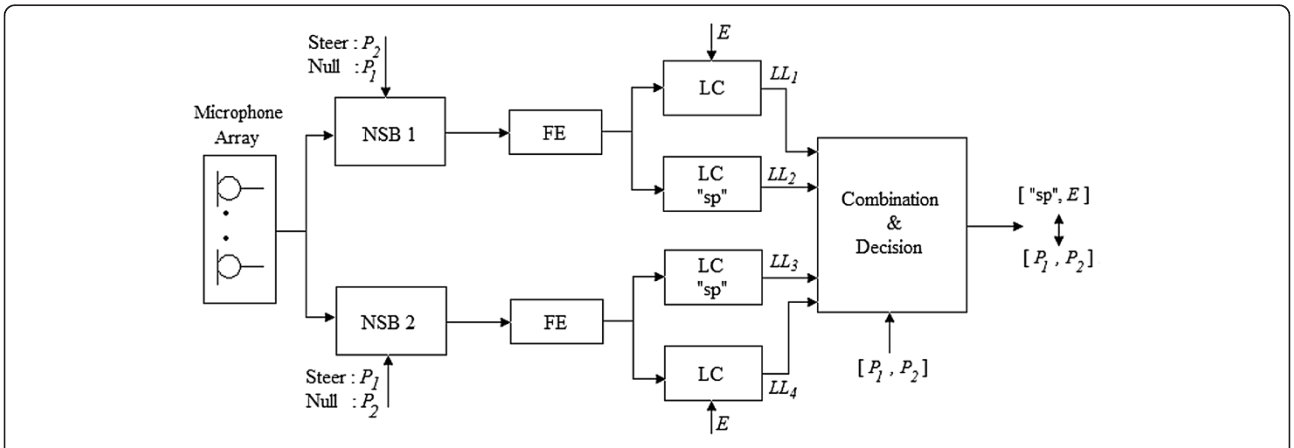

Figure 3 Position assignment system for a single array.

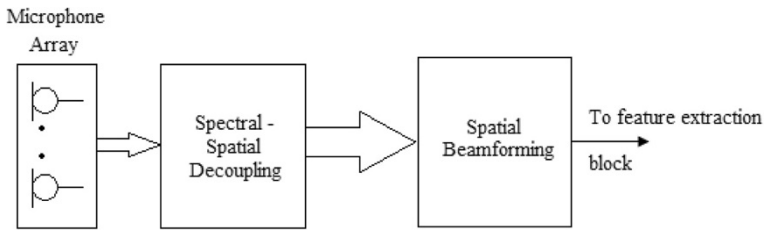

Figure 4 Frequency-invariant beamforming.

localization property of the LFBEs. In the experiments, a 32-dimension feature vector (16 FF-LFBEs and their first temporal derivatives) is used. As shown in the scheme of Figure 3, there is a set of two likelihood calculators for each parallel channel, one of them to calculate the model based log-likelihood for the AE label (E), provided by the AED system, and the other to calculate it for speech. As in [14], here also hidden Markov models (HMMs) are employed, where Gaussian mixture models (GMMs) are used to calculate the emission probabilities [29]. Thirty-two Gaussian components with diagonal covariance matrices are used per model. There is one left-to-right HMM with three emitting states for each AE and speech. Eleven HMMs are trained with isolated events using the Baum-Welch algorithm [29]. The HTK toolkit is used for training and testing this HMM-GMM system [30].

To make the mapping between source positions and event identities, the decision block uses the four log-likelihood scores computed from the HMM-GMM models. Those four scores, which are indicated in Figure 3 as LL_i, $i = 1, 2, 3, 4$, are grouped in two log-likelihood ratios LLR_1 and LLR_2, one for each beamformer path, and the following single-array score S is computed:

$$S = LLR_1 + LLR_2 = (LL_1 - LL_2) + (LL_3 - LL_4) \qquad (1)$$

If S is positive, the AE E is associated to the position P_2, and if S is negative, it is associated to P_1. Let us illustrate it with a particular case. Assume that P_1 truly corresponds to speech and P_2 to the acoustic event E. When using the AE model, it is expected to get comparatively higher log-likelihood from the output of NSB1 (LL_1) than from the output of NSB2 (LL_4). For the clean speech model, it is expected to get comparatively higher log-likelihood from the output of NSB2 (LL_3) than from the output of NSB1 (LL_2). If that is the case, the decision is taken that speech is at P_1 and E is at P_2, which is the correct decision. Note that with this type of combination, the decision block gives equal importance to all the four likelihood calculator outputs.

In order to get the most from the available information, in the current scheme, unlike in [15], the classification stage includes a speech model besides the AE model. Indeed, the system could also work with only either the speech model-based classifier or the AE model-based classifier. To study the contribution of each one of the models, all those options have been tested, and the results are reported in Section 4. For the decision, if only either the AE or the speech-based classifier is used, just either LL_1-LL_4 or LL_3-LL_2, respectively, is needed.

3.5. Multi-array fusion

As it is mentioned earlier, all the six three-microphone linear arrays deployed in the room are used in the position assignment system. For taking the assignment decision, the six sets of scores LLR_1 and LLR_2, computed as indicated in Equation 1, are combined either with a uniformly weighted average [31] of the 12 values or by fuzzy integral-based fusion [19]. In the following, the latter technique is presented.

3.5.1. FI-based optimized fusion

The scores at the output of the classification stage can be linearly combined by using an optimal fusion approach that assigns an individual weight to each of them. However, a more sophisticated weighting technique that considers all subsets of information sources: the fuzzy integral (FI) approach, is considered in this work [19].

Let us denote by h_i, $i = 1, 2,..., N$, the set of output scores (LLR_1 and LLR_2) of the $N/2$ single-array systems. Assuming that the sequence h_i, $i = 1, 2,..., N$, is ordered in such a way that $h_1 \le ... \le h_N$, the Choquet fuzzy integral can be computed as

$$M_{\mathrm{FI}}(\mu, h) = \sum_{i=1}^{N} [\mu(i, ..., N) - \mu(i+1, ..., N)] h_i \qquad (2)$$

where $\mu(N+1) = 0$. The value $\mu(Q)$ can be viewed as a weight related to a subset Q of the set Z of N information sources. It is called the fuzzy measure, and if Q and T are subsets of Z, it has to meet the following conditions:

Boundary: $\mu(\phi) = 0, \mu(Z) = 1$

Monotonicity: $Q \subseteq T \Rightarrow \mu(Q) \le \mu(T)$

In this work, a supervised gradient-based training algorithm for learning the fuzzy measures from the training data with cross-validation is used [18,19].

4. Experimental work

The PA experiments are done under the assumption that there is always an AE overlapped with speech. It is assumed that the identity of the AE event is known, to avoid the propagation of the AED errors to the PA system. Additionally, it is assumed that the approximate position in the room of the AE source and the speaker are known. Thus, the PA system only has to make a binary decision (the AE is from position P_1 or position P_2), which will be either correct or incorrect.

To design and evaluate the performance of the system, the position assignment rate (PAR) metric for a given AE class is defined as the quotient between the number of correct decisions and the total number of occurrences of that class in the testing database. Then, the PAR will be averaged over the classes to have the final evaluation measure. For reference, a second metric is also considered, called Diff_LL, which is the value of the S score from Equation 1 provided that the assignment is correct (LL_1-LL_4 for the AE-based system, or LL_3-LL_2 for the speech-based one, or S when both the AE model and the speech model are used). Actually, that score can be considered as an estimate of the degree of source separation carried out by the beamformers when a correct assignment is made.

In the PA system from Figure 3, there exist two FIB-based NSBs at the front end. The design of the beamformers for each particular AE sample requires the DOA angles corresponding to the target and the null, i.e., the DOAs from the source positions P_1 and P_2. Two different options regarding the approximate positions of the acoustic events from which the DOAs are extracted have been considered. First, the same approximate DOA for the whole set of acoustic events for each array has been considered. It is obtained as a DOA average over the AE source positions, which are known from visual inspection during recording. In this case, a beam pattern with a broader main lobe (as shown in the right side of Figure 5) to approximately encompass all the positions of the acoustic events has been designed. And, in the second option, the position of the event estimated by an ASL system based on the SRP-PHAT technique has been used. Therefore, in that case, the beam steers to the direction of the specific event position. It is worth to mention here that the AE source positions are estimated using a one-source ASL system instead of a two-source one, in order to avoid more propagation of errors from the ASL system to the PA system. Regarding the speech source position, the speaker's position specified during recording has been used for all experiments.

4.1. Results and discussion

To assess the performance of the PA system depicted in Figure 3, several experiments have been conducted. The testing results are obtained with all the eight recording sessions (S01 to S08), using a leave-one-out criterion, and averaging over all the testing dataset. In all the FI-based fusions, a 5-fold cross-validation on the training data to stop the training process and avoid over-fitting is used. To check the performance of the PA system when either only the AE model or only the speech model is used, the experiments for the array T6 is performed, using visually inspected positions for AEs and a broad beam. The results are shown in Table 2. It is worth mentioning that the AE source positions and the speech source position are physically rather well separated from the viewpoint of the array T6.

It can be observed, from the results in Table 2, that the combination of the two models with the S score, which averages the scores LLR_1 and LLR_2, improves the

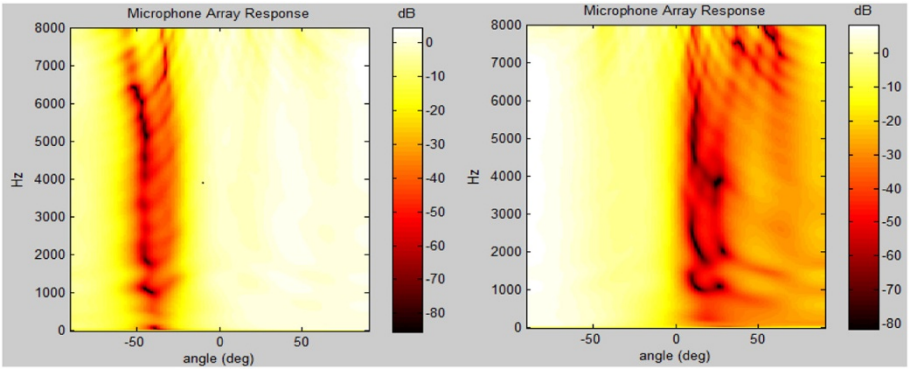

Figure 5 Beam pattern of the FIB. Left: null towards speech; right: null towards the AEs in the case of a wide beam encompassing the DOAs of all the AEs.

Table 2 PA rate and Diff_LL for the PA system with the T6 array alone

	AE model	Speech model	Average of LLR scores	FI-based fusion
PAR (%)	86.2	81	87.1	91.2
Diff_LL	1.44	1.07	2.53	2.88

performance of the system with respect to the use of only one type of model. The improvement is much more noticeable using the FI-based fusion of the two scores. Notice also that the AE model-based system works much better than the speech model-based one. In fact, the former uses a more specific model, because the speech model is obtained from the whole set of speech sounds. In that table, the Diff_LL score is also shown. Notice that, in general, it is well correlated with the PAR one. However, there is a large difference between the values of Diff_LL for the AE-based case and the LL combination case, in contrast with the very small difference there is in terms of PAR. It means that the use of both models allows achieving a much stronger confidence on the PA decision when it is correct.

Table 3 shows the PAR scores when all six microphone arrays are employed, either alone or in combination. The results are given for the two types of DOA settings mentioned above. Also, two types of intra-array combinations are considered, as in Table 2 (second column of Table 3): the average of LLR scores given by Equation 1 and the FI-based fusion. The testing results are obtained with all the eight recording sessions (S01 to S08), using a leave-one-out criterion, averaging over all the testing dataset, and tabulated with the population standard deviations.

In the first columns of Table 3, we have the PAR (in %) with the standard deviations when each one of the arrays is used alone. Notice that the PAR scores of the upper half of the array numbers (T4, T5, and T6) are higher than the ones of the lower half. It could be expected, since for those arrays the angle between the acoustic event positions and the speech source position is larger than that for the other arrays (T1, T2, and T3).

Note from the two last columns in Table 3 that the accuracy obtained from either the average of LLR scores or the FI-based fusion of the whole set of arrays is higher than the accuracy obtained from any of the single arrays. Comparing both types of fusion, the FI one shows a noticeable better performance for both DOA setting cases, arriving to a PA error of only 4.3%.

The use of intra-array FI-based fusion improves the PAR scores with respect to using a uniformly weighted average of LLR scores, especially for the upper half arrays. Therefore, though by employing the FI approach there is the cost of having to learn the fuzzy measures from data, it may be a good choice when the quality of the signal separation is not too low, like it presumably happens with the upper half arrays

Regarding the type of DOA setting, the ASL-estimated AE position-based system works always better than the one that uses an average DOA based on visual inspection (i.e., a broad-beam nulling angle). That could be expected, since the beam pattern is specific of each event occurrence, whereas the broad beam encompasses all the angles of the AE source positions. While the latter design simplifies the overall system, as it does not require a precise source position and may avoid an additional external ASL block, it is specific of the given scenario, so it has to be redesigned when the scenario changes.

5. Conclusions

An attempt is made in this paper to resolve the source identification ambiguity that appears when an acoustic event, which overlaps with speech, is detected. A position assignment system has been proposed and tested. It firstly consists of a set of frequency-invariant null-steering beamformers that carry out different partial signal separations for each microphone array. The beamformers are followed by model-based likelihood calculations, using both the acoustic event model and the speech model, to obtain two likelihood ratios, whose combination gives a final score per array. Using the fuzzy integral for that intra-array combination and also for the fusion of the six array scores, the best assignment error is obtained, which is smaller than 5%. It is worth noticing that, though the position assignment system has been developed for the problem encountered in the current scenario, its scheme can be extended to more than two sources and to different types of sound overlap combinations. Future work will be devoted to that.

Table 3 PA performance (%) with standard deviation for each single array and for the two combinations

DOA setting	Intra-array combination	T1	T2	T3	T4	T5	T6	Inter-array combination	
								Average of LLR scores	FI-based fusion
Broad-beam nulling angle	Average of LLR scores	83.1 ± 1.9	77.3 ± 2.6	81.3 ± 1.9	88.9 ± 2	88.2 ± 1.8	87.1 ± 1.8	89.8 ± 2	93.5 ± 1.9
	FI-based fusion	83.5 ± 1.9	77.1 ± 2.5	82.3 ± 1.8	92.8 ± 1.4	92.7 ± 1.5	91.2 ± 1.6	-	93.6 ± 1.7
ASL-estimated AE positions	Average of LLR scores	88.2 ± 1.8	85.6 ± 2.3	89.8 ± 1.7	91.2 ± 1.6	92.1 ± 1.7	91 ± 1.9	93.6 ± 1.8	95.4 ± 1.7
	FI-based fusion	88.3 ± 1.7	84.9 ± 2.1	90.2 ± 1.6	92.7 ± 1.3	93 ± 1.4	92.2 ± 1.4	-	95.7 ± 1.5

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work has been supported by the Spanish project SARAI (TEC2010-21040-C02-01).

Received: 4 August 2013 Accepted: 21 March 2014
Published: 28 April 2014

References

1. W Wang, *Machine Audition: Principles, Algorithms and Systems* (IGI Global, 2010)
2. A Waibel, R Stiefelhagen, *Computers in Human Interaction Loop* (Springer, New York, 2009)
3. A Temko, C Nadeu, D Macho, R Malkin, C Zieger, M Omologo, Acoustic event detection and classification, in *Computers in the Human Interaction Loop*, ed. by A Waibel, R Stiefelhagen (Springer, New York, 2009), pp. 61–73
4. X Zhuang, X Zhou, MA Hasegawa-Johnson, TS Huang, Real-world acoustic event detection. Pattern Recogn. Lett. **31**, 1543–1551 (2010)
5. M Omologo, P Svaizer, Acoustic event localization using crosspower-spectrum phase based technique, in *ICASSP* (Adelaide, 1994)
6. T Nishiura, S Nakamura, Study of environmental sound source identification based on hidden Markov model for robust speech recognition. J. Acoust. Soc. Am. **114**(4), 2399 (2003)
7. H Wang, P Chu, Voice source localization for automatic camera pointing system in videoconferencing, in *ICASSP* (Munich, 1997), pp. 187–190
8. J DiBiase, HF Silverman, M Brandstein, Microphone arrays. Robust localization in reverberant rooms, in *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M Brandstein, D Ward (Springer, New York, 2001)
9. D Wang, GJ Brown, in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, ed. by D Wang, GJ Brown (Wiley-IEEE, 2006)
10. J Dmochowski, J Benesty, Steered beamforming approaches for acoustic source localization, in *Speech Processing in Modern Communication*, ed. by I Cohen, J Benesty, S Gannot, vol. 12 (Springer, Berlin, 2010), pp. 307–337
11. J Velasco, D Pizarro, J Macias-Guarasa, Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints. Sensors **12**(10), 13781–13812 (2012)
12. CLEAR, *Classifications of Events, Activities and Relationships: Evaluation and Workshop* (Baltimore, 2007)
13. A Temko, C Nadeu, Acoustic event detection in meeting-room environments. Pattern Recogn. Lett. **30/14**, 1281–1288 (2009)
14. T Butko, F Gonzalez Pla, C Segura, C Nadeu, J Hernando, Two-source acoustic event detection and localization: online implementation in a smart-room, in *Proc. EUSIPCO* (Barcelona, 2011)
15. R Chakraborty, C Nadeu, T Butko, Detection and positioning of overlapped sounds in a room environment, in *Proc. Interspeech* (Portland, 2012)
16. R Chakraborty, C Nadeu, T Butko, Binary position assignment of two known simultaneous acoustic sources, in *Proc. IberSPEECH* (Madrid, 2012)
17. M Grabisch, Fuzzy integral in multi-criteria decision-making. Fuzzy Set Syst **69**(3), 279–298 (1995)
18. S Chang, S Greenberg, Syllable-proximity evaluation in automatic speech recognition using fuzzy measures and a fuzzy integral, in *Proc. IEEE Fuzzy Systems Conference* (St. Louis, 2003), pp. 828–833
19. A Temko, D Macho, C Nadeu, Fuzzy integral based information fusion for classification of highly confusable non-speech sounds. Pattern Recogn. **41**(5), 1831–1840 (2008)
20. BD Van Veen, KM Buckley, Beamforming: a versatile approach to spatial filtering. ASSP Mag IEEE **5**(2), 4–24 (1988)
21. SP Applebaum, Adaptive arrays. IEEE Trans Antenna Propag **24**, 585–595 (1976)
22. AS Feng, DL Jones, Localization-based grouping, in *Computational Auditory Scene Analysis: Principles, Algorithm, and Applications*, ed. by D Wang, GJ Brown (IEEE/Wiley-Interscience, 2006)
23. O Hoshuyama, A Sugiyama, Robust adaptive beamforming, in *Microphone Arrays: Signal Processing Techniques and Applications*, ed. by M Brandstein, D Ward (Springer, New York, 2001)
24. DB Ward, RA Kennedy, RC Williamson, Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns. J. Acoust. Soc. Am. **97**, 1023–1034 (1995)
25. LC Parra, Least squares frequency invariant beamforming, in *Proc. Workshops on Applications of Signal Processing on Acoustics and Audio* (IEEE, New York, 2005)
26. LC Parra, Steerable frequency-invariant beamforming for arbitrary arrays. J. Acoust. Soc. Am. **119**(6), 3839–3847 (2006)
27. Y Zhao, W Liu, RJ Langley, Design of frequency invariant beamformers in subbands, in *Proc. IEEE/SP 15th Workshop on Statistical Signal Processing* (Cardiff, 2009)
28. C Nadeu, D Macho, J Hernando, Frequency and time filtering of filter-bank energies for robust HMM speech recognition. Speech Comm. **34**, 93–114 (2001)
29. L Rabiner, B Juang, *Fundamentals of Speech Recognition* (Prentice Hall, 1993)
30. S Young, G Evermann, D Kershaw, G Moore, J Odell, D Ollason, D Povey, V Valtchev, P Woodland, *The HTK Book (for HTK Version 3.2)* (Cambridge University, 2002)
31. LI Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms* (Wiley-Interscience, New Jersey, 2004)

doi:10.1186/1687-4722-2014-18
Cite this article as: Chakraborty *et al.*: **Source ambiguity resolution of overlapped sounds in a multi-microphone room environment.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:18.