Model-Based Processing for Acoustic Scene Analysis

Climent Nadeu, Rupayan Chakraborty and Martin Wolf

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

ABSTRACT

The analysis of acoustic scenes requires several functionalities, being perhaps recognition (speech, speaker, other acoustic events) and spatial localization the two most relevant ones. For a reduced invasiveness, the microphones are far away from the sound sources, and possibly grouped in arrays, which may be distributed, not arranged, in the room. Aiming at an increased performance, the usual model-based approach employed for sound recognition or detection can be extended to other co-occurrent tasks like source localization, so both tasks can be carried out jointly, using the same formulation and processing. In this paper, we intend to illustrate that point by presenting together a few new model-based techniques that deal with the problems of overlapped-sounds recognition, multi-source localization, and channel selection. They are briefly described, and tested in a smart-room environment with a multiple microphone-array setup.

Index Terms— Acoustic scene analysis, audio recognition, acoustic source localization, channel selection, multi-microphone processing.

1 INTRODUCTION

The aim of acoustic scene analysis (ASA) is to describe the sequences of entities that are conveyed by the acoustic signals produced in a given environment, as well as to determine the time positions of those entities and the spatial locations of their sources. Therefore, apart from the preparatory processing (noise removal, de-reverberation, source separation, etc.), two basic functionalities are required for ASA: 1) recognition (speech, other audio events, speaker identity,…) or detection, which includes estimation of the position of the sounds along the time dimension; and 2) spatial localization of the sound sources. To reduce invasiveness, the microphones are placed on the walls, tables, etc. so they may be distant from the sources, and often grouped in arrays. For cost effectiveness and deployment facility, requirements which are usual in an environment like the domestic one, it is preferable that the arrays: 1) can be distributed, not arranged in space, 2) can be placed in different rooms, and 3) can consist of a small number of microphones.

On the one hand, the recognition (or detection) functionality is an usual requirement in ASA. On the other hand, current state-of-the-art systems for audio recognition are mostly based on models, meaning that they use some a-priori learned representation of the sound or the sound elements. Consequently, a set of sound models is usually available for other ASA tasks, e.g. source localization.

Indeed, models are also used in the pre-processing steps of acoustic scene analysis; e.g. for source segregation [1]. Recently, a model-based approach was used for selecting the microphones whose signals are the least affected by reverberation so they may yield the highest speech recognition accuracy [2]. That kind of selection can be used before any other processing, either if arrays are used or not.

Moreover, despite most of the existing localization algorithms are blind with respect to the content of the signal, the information provided by a set of sound models can be used advantageously to improve the localization results, as it was recently shown [3].

In this paper we intend to emphasize the interest of the model-based approach for the various ASA tasks other than the recognition one, by presenting together different techniques we have been working recently. In particular, both sound source localization and channel selection are considered in combination with audio recognition. Interestingly enough, the model-based approach allows that, by using a common multi-channel processing scheme and a shared set of statistical models, both localization and recognition of multiple simultaneous sound sources can be jointly carried out with a MAP formulation.

A selection of specific techniques and experimental results is presented to illustrate that point in a particular scenario consisting of either meeting-room acoustic events overlapped with speech or reverberated digit strings, which were recorded in a room with 6 arrays of 3 microphones hanging on the walls.

2 RECOGNITION AND LOCALIZATION OF SIMULTANEOUS ACOUSTIC EVENTS

Acoustic event recognition/detection may be a challenging task. From the analysis of the results submitted by the

participants in the meeting-room acoustic event detection (AED) task of the CLEAR'07 evaluation campaign [4], it was apparent that time overlapping of acoustic events caused more than two thirds of the detection errors. More recently, in the D-CASE evaluation [5] a big gap was observed between the AED accuracy of isolated and overlapped events. In terms of F-score, the best results were 45.17% and 8.45%, respectively.

The detection of simultaneous acoustic events may be dealt with different approaches. First, at the model level, e.g. by training a model for each different sound combination [6]. Second, at the feature level, e.g. using features from other modalities, like video features or audio localization features [7]. The third approach we have considered [8-9], combines beamforming-based source separation with signal modeling using a MAP criterion.

Most current acoustic source localization (ASL) techniques rely on energy-like measures extracted from the microphone signals. Conversely, in [3], the use of information about the content of the signals is proposed. In fact, instead of relying only on signal-based measures, the probability measure delivered by a classifier that uses models for the different sound classes is proposed. Furthermore, since not only the sound models but also the processing scheme are shared with the AED system, both tasks, localization and recognition, can be carried out jointly. That joint approach [9], already introduced in [10] in terms of estimation of the angle of arrival (not position) is presented in the next sub-sections.

2.1 Multi-microphone processing scheme

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. For each microphone array, there is a set of P beamformers, each one attenuating the signals from all directions except the direction corresponding to a 2-D position in the room s_j, $1 \le j \le P$ (this development could be easily extended to 3-D).

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, by using previously trained acoustic event models. Finally, a decision module carries out the localization of the events by combining the likelihood scores with a MAP criterion.

A null-steering beamformer (NSB) is capable of placing nulls at different positions in the sensor array patterns. Given the broadband characteristics of the audio signals, in order to determine the beamformer coefficients we use a technique called frequency invariant beamforming. The method, proposed in [11], 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. The 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.

2.2 MAP-based joint classification and localization

Given a room with K microphone arrays, let us assume we have a set of N (possibly simultaneous) events c_i, $1 \le i \le N$, which belong to a set of C 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 beamformers, so that the j-th beamformer is attenuating the signals coming from the directions corresponding to all positions, except that of position s_j. So from array processing, we have a set of P output signals for each array, and after likelihood computations with the models of all classes, we have a PxCxK-dimensional vector of likelihood scores.

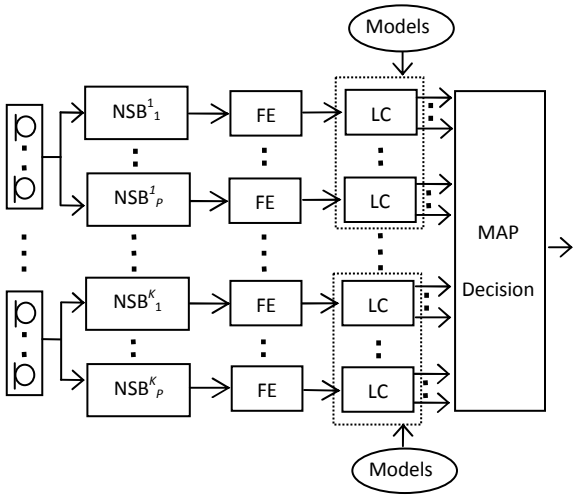

Figure 1. Scheme for both classification and localization.

We want to determine the posterior probability of a given class c_i and position s_j for the k-th array,

$$p(c_i, s_j \mid X_k) = p(X_k \mid c_i, s_j) p(c_i) p(s_j) / p(X_k) \qquad (1)$$

where X_k denotes the multi-channel signal from the k-th array. For combining the posterior probabilities from the various microphone arrays, we will use the product combination rule, so the optimal class c_o and the optimal position s_o is chosen to maximize a product of posterior probabilities [12], i.e.

$$c_o, s_o = \underset{c_i, s_j}{\text{argmax}} \prod_{k=1}^{K} p(c_i, s_j | X_k) \qquad (2)$$

In the case of N simultaneous sources, and assuming they correspond to N different classes, the recognized

identities of those classes and the corresponding positions are obtained by applying Eq. (2) N consecutive times and leaving each time the recognized class and its corresponding position out.

2.3 Experimental work

In our experimental work, we consider a meeting room scenario with a predefined set of 11 acoustic events plus speech [4].

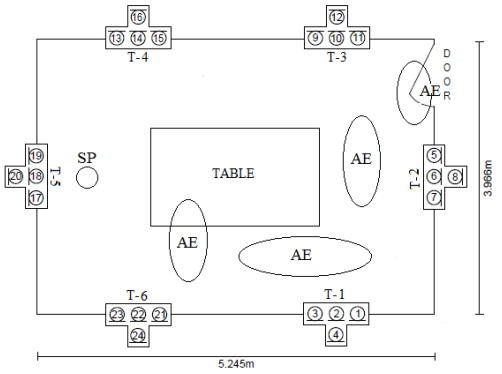

Figure 2. Smart-room layout, with positions of microphone arrays (T-*i*), acoustic events (AE) and speaker (SP)

Fig. 2 shows the smart-room, with the position of its six T-shaped 4-microphone arrays on the walls. Only the linear arrays of 3 microphones are used in the experiments. For training, development and testing of the system, we have used part of a publicly available multimodal database recorded in the smart-room [7]. Concretely, 8 recording sessions of audio data are used, which contain isolated acoustic events. The approximate source positions of the acoustic events (AE) are shown in Fig. 2. Each session was recorded with all the six T-shaped microphone arrays.

2.3.1 Acoustic event classification and localization

The experiments reported herewith correspond to the case of having a single acoustic event overlapped with speech. All signals were recorded at 44.1 kHz sampling frequency, and further converted to 16 kHz. Additionally, in these experiments we assume the AE time endpoints are known, so the determination of their identities results in a classification problem. That assumption can be removed and so the problem extended to detection by using the same processing scheme from Section 2.1 [13]. 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). In the reported experiments, the proposed system depicted in Fig. 1 is used both to recognize and to localize the acoustic event sources. For that purpose, the 2-D room space is divided into a set of P pre-defined cells. To facilitate real time processing, a relatively large cell size has been considered: 0.6x0.8m.

In the feature extraction block of the system, 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. Frequency-filtered log filter-bank energies (FF-LFBE) are used for the parametric representation of the spectral envelope of the audio signal [14]. 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 [15]. 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. For each array, the likelihoods are computed by using the same set of acoustic event and silence models for all the beamformer outputs. The optimal class identity and source position are obtained according to Eq. (2). Instead of using the explicit prior probabilities from Eq. (1), in the reported tests, a machine learning based non-linear transformation technique was used to convert the likelihoods into posterior probabilities before applying the MAP criterion, like in [10].

2.3.2 Experimental results

Testing results with all the 8 sessions (S01-S08) are obtained using all the six arrays (T1 to T6) available in the room. A leave-one-out criterion is used, i.e. recursively keeping one session for testing, while all the other 7 sessions are used for training. Table 1 shows the classification error for the proposed system (right-hand side). For comparison, the error rate from two other systems is shown: the system that assumes the source positions are known, and also the above mentioned reference system that uses a single microphone (no beamforming) and a model for each different sound combination [6].

Notice that the proposed system, which jointly estimates the AE identity (the other event is always speech), and its cell-based position, gets a lower error rate than our reference AED system, though a higher one than the (virtual) system that assumes the source positions are known.

The localization results regarding the AE source location are also shown in the second row of Table 1 in terms of cell error (number of events assigned to the correct room cell divided by total number of events). For comparison, the errors from the system that assumes knowledge of the event identity are also presented, together with the ones from the reference SRP-PHAT system [16-17]. All systems perform localization at the event level.

As observed in Table 1, though the proposed method would perform better if the sound identities were known, it

clearly outperforms the SRP-PHAT system, which is implemented by looking at the two maxima of the sound map. It is worth noticing that, for the proposed system, the cell error rate is only 2.2% higher than that of the one-source case presented in [3] (13.4%).

	Reference systems	Systems that assume position or identity	Joint classif. and localization system
Classif. error (%)	16.2	10.9	12.4
Localiz. error (%)	29.1	14.5	15.6

Table 1. Percentages of classification error and cell localization error for the various tested systems.

3 CHANNEL SELECTION

In applications where distant-talking microphones are used for automatic speech recognition (ASR), both additive noise and room reverberation are major factors of recognition rate degradation [18]. If multiple microphones are available, signal combination techniques (e.g. beamforming) are often used to improve the quality of the acquired speech. However, this combination may not be possible, or the quality of the combined signal may not be always better than the quality of the signal from the single best channel.

In a typical domestic scenario, the microphones may be arbitrarily located and have different characteristics: a living room where some microphones are hanging on the walls, others standing on a table or shell, and others are built in the personal communication devices of the people. In such situation, the ASR performance may actually gain from neglecting the most disturbed channels (microphone level), or group of channels (array level), or even from selecting just the hypothesized best channel for recognition. So a channel ranking is needed. Ideally, the single selected channel would be the one that leads to the lowest word error rate (WER) after recognition. Since WER is unknown during the recognition process, a different measure, as correlated as possible with the WER is needed.

Several channel selection (CS) techniques were proposed in [2,19], either using a model-based (or decoder-based) approach or a signal-based one, and they were evaluated together with other measures which had already been proposed in the literature. The model-based measure which is considered here is computed from the *N*-best hypothesis list at the output of the ASR system [20].

Starting from the Bayes rule and substituting the probability of the observation in the m-th channel $p(O_m)$ by its approximation from the *N*-best hypothesis, the following measure of quality C_m is obtained for channel m:

$$C_m = \frac{p\left(O_m \middle| w_m^1\right)^{1/\alpha} p\left(w_m^1\right)}{\sum_{n=1}^{N} p\left(O_m \middle| w_m^n\right)^{1/\alpha} p\left(w_m^n\right)} \quad (3)$$

where w_m^n is the n-th word sequence (out of the list of N) hypothesized for channel m. A parameter α is included in the exponent, which is made equal to the number of frames.

3.1 Experimental work with digit recognition

For experimentation, the signals of the English TIDigits database were convolved with impulse responses measured in the smart-room depicted in Figure 2. This set of room impulse responses (which is freely available from the authors) consists of 7 source positions, spread in the room, and, for each one, 4 orientations were considered (for the loudspeaker). It amounts 28 impulse responses per each one of the 24 microphones.

Table 2 shows the results obtained from an experiment with reverberated digit strings [20]. The above mentioned model-based technique is compared with an effective signal-based technique that relies on a measure of variance of the time envelope (i.e. a kind of modulation index) of the speech signal, which is affected by the room reverberation [2]. Both clean speech training and matched training are tested.

CS method	*Clean training*	*Matched training*
Random choice	22.7	6.0
Envelope variance	19.2 (15.4%)	5.5 (8.3%)
N-best	17.6 (22.5%)	4.9 (18.3%)

Table 2. WER (in %). The relative improvement with respect to the random choice is shown in parenthesis.

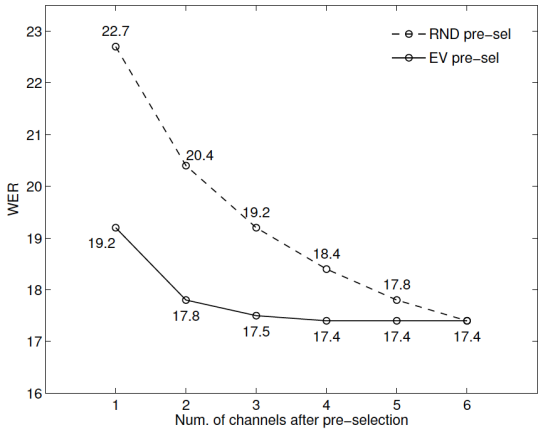

Figure 3. WER after doing channel selection in two steps. EV refers to the envelope variance CS method.

From Table 2, we can observe the advantage of using a CS technique with respect to using a random choice of the microphone for each digit string. Results for other types of CS techniques are reported in [2]. Notice that the model-based approach shows better accuracies than the signal-based one. However, it implies a higher complexity. There is a way to combine the advantages of both, by doing a pre-selection with the simpler signal-based technique, and a final selection with the model-based one.

Figure 3 shows the CS performance for the clean speech training case in terms of the number of pre-selected channels using a pre-selection that is either random or based on the above mentioned envelope variance (EV) CS method [2]. The costly model-based measure has to be computed only for 3 channels without almost no loss with respect to using all 6 channels. In the second selection step, both measures are used in combination, as explained in [20].

4 CONCLUSION

Though in the reported work the experiments concerning recognition and localization have been carried out with acoustic events, and the experiments about channel selection have been presented in the context of speech (digit) recognition, all the reported techniques can be applied to any sound classification task. Furthermore, the techniques may be combined and the models shared by all the systems; for instance, a channel selection technique could be used to weight the contribution of each microphone array in Eq. (2).

5 ACKNOWLEDGMENTS

The authors would like to thank Andrey Temko and Taras Butko, who made the initial steps in the work presented here about AED.

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

6 REFERENCES

[1] W. Wang, Ed., *Machine Audition: Principles, Algorithms and Systems*, IGI Global Press, 2010.

[2] M. Wolf and C. Nadeu, “Channel selection measures for multi-microphone speech recognition,” *Speech Communication*, vol. 57, pp. 170–180, Feb. 2014.

[3] R. Chakraborty and C. Nadeu, “Sound-model-based acoustic source localization using distributed microphone arrays”, in *Proc. ICASSP*, Florence, Italy, 2014.

[4] 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.

[5] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. Plumbley, “Detection and classification of acoustic scenes and events: an IEEE AASP challenge”, in *Proc. WASPAA*, 2013.

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

[7] T. Butko, C. Canton-Ferrer, C. Segura, X. Giró, C. Nadeu, J. Hernando, and J. R. Casas "Acoustic event detection based on feature-level fusion of audio and video modalities", *EURASIP Journal on Advances in Signal Processing*, Vol. 2011, Article ID 485738, 2011.

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

[9] R. Chakraborty, *Acoustic source detection and localization using distributed microphone arrays*, PhD thesis dissertation, Dec. 2013.

[10] R. Chakraborty and C. Nadeu, "Joint recognition and direction-of-arrival estimation of simultaneous meeting-room acoustic events", *in Proc. Interspeech*, Lyon, France, Aug. 2013, pp. 2948-2952.

[11] L.C. Parra, “Steerable Frequen*cy-Invariant Beam*forming for *Arbitrary A*rrays”, *Journal of the Acoustical Society of America*, 119 (6), pp. 3839-3847, June, 2006.

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

[13] R. Chakraborty and C. Nadeu, "Model-based joint detection and localization of acoustic events using distributed microphone arrays”, in preparation.

[14] 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.

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

[16] M. Omologo and P. Svaizer, "Use of crosspower-spectrum phase in acoustic event detection," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 5, no. 3, pp. 288-292, 1997.

[17] 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,” *in Proc. ICASSP*, Hawaii, USA, 2007.

[18] M. Wölfel, J. McDonough, *Distant Speech Recognition*, Wiley Hoboken, NJ, 2009.

[19] M. Wolf, *Channel selection and reverberation-robust automatic speech recognition*, PhD thesis dissertation, Nov. 2013.

[20] M. Wolf and C. Nadeu, "Channel selection using N-best hypothesis for multi-microphone ASR", *in Proc. Interspeech,* Lyon, France, Aug. 2013, pp. 3507-3511.