MULTI-MICROPHONE FUSION FOR DETECTION OF SPEECH AND ACOUSTIC EVENTS IN SMART SPACES

Panagiotis Giannoulis[1,3], *Gerasimos Potamianos*[2,3], *Athanasios Katsamanis*[1,3], *Petros Maragos*[1,3]

[1] School of Electr. and Computer Eng., National Technical University of Athens, 15773 Athens, Greece
[2] Department of Electr. and Computer Eng., University of Thessaly, 38221 Volos, Greece
[3] Athena Research and Innovation Center, 15125 Maroussi, Greece

gpotam@ieee.org, nkatsam@cs.ntua.gr, maragos@cs.ntua.gr

ABSTRACT

In this paper, we examine the challenging problem of detecting acoustic events and voice activity in smart indoors environments, equipped with multiple microphones. In particular, we focus on channel combination strategies, aiming to take advantage of the multiple microphones installed in the smart space, capturing the potentially noisy acoustic scene from the far-field. We propose various such approaches that can be formulated as fusion at the signal, feature, or at the decision level, as well as combinations of the above, also including multi-channel training. We apply our methods on two multi-microphone databases: (a) one recorded inside a small meeting room, containing twelve classes of isolated acoustic events; and (b) a speech corpus containing interfering noise sources, simulated inside a smart home with multiple rooms. Our multi-channel approaches demonstrate significant improvements, reaching relative error reductions over a single-channel baseline of 9.3% and 44.8% in the two datasets, respectively.

***Index Terms*—** acoustic event detection and classification, voice activity detection, multi-channel fusion

1. INTRODUCTION

Acoustic event detection (AED) constitutes a research area that has been increasingly gaining interest. Among others, its application to smart space environments, such as homes or offices equipped with multiple sensors including microphone arrays, can reveal valuable information about human and other activity, which can be useful to the development of smart space applications. Moreover, detection of acoustic events can also improve performance of core speech technologies, such as automatic speech recognition (ASR) and enhancement. AED, in general, aims to identify both time boundaries and the type of the event(s) occurring.

Various AED approaches have been proposed in the literature, varying in the features employed and the detection and classification methods used [1–4]. However, most operate on single-microphone audio input, with only a few exploiting information from multiple microphones. Among the latter, in [5], outputs of support vector machines from each channel are combined via majority voting for AED. Also, in [6], channel decisions are firstly fused by averaging their log-likelihood scores at each frame, and then an optimal path of events is computed by the Viterbi decoding algorithm. Finally, in [7], decisions from different modalities are fused employing a fuzzy integral statistical approach to cope with the problem of overlapping event detection.

Related to the above is the problem of voice activity detection (VAD) that can be viewed as a special AED case with two only classes of interest (speech, non-speech). VAD has attracted significant research interest, due to its importance to ASR and human-computer interaction. Among others, developed single-channel VAD systems employ energy thresholding [8], statistical modeling [9], and discriminative frontends [10]. Concerning multi-microphone approaches, in [11], majority voting is used to fuse single-channel VAD outputs, while, in [12], homogeneity of time-delays between two microphone signals is exploited.

In our work, we address both AED and VAD problems within a single framework, focusing on multi-channel approaches to exploit information from the available microphones at various levels. In particular, we investigate channel fusion at the signal level, employing beamforming techniques to produce enhanced signals, at the feature level, utilizing time-difference-of-arrival (TDOA) between channel signals as additional informative features, and at the decision level, appropriately integrating detection decisions to yield the final one. Further, "multi-style" training is also considered, utilizing observations from all available microphones to produce more robust models.

The above are investigated using two related detection systems that are based on appropriately trained Gaussian mixture models (GMMs) on traditional audio front-end features. The first is a frame-based GMM that operates over sliding windows of fixed duration, whereas the second employs Viterbi decoding over the entire observation sequence,

This research was supported by the EU project DIRHA with grant FP7-ICT-2011-7-288121.

based on a hidden Markov model (HMM) composed of the trained GMMs over the classes of interest. Experimental results are reported on two multi-microphone corpora, one containing isolated acoustic events of twelve types occurring in a single room that is appropriate for AED, and a second one containing speech and interfering noise simulated inside a multi-room apartment, appropriate for VAD. In both cases, multi-channel approaches are demonstrated to significantly outperform single-channel baselines.

The rest of the paper is organized as follows: Section 2 presents the multi-channel methods for fusion and information extraction; Section 3 describes details of the two detection approaches used; Section 4 is devoted to the experiments and results; and, finally, Section 5 concludes the paper.

2. MULTI-CHANNEL INFORMATION EXTRACTION AND FUSION

A number of channel combination approaches at different levels are investigated in this paper, as discussed next.

2.1. Multi-channel training

In this approach, observations from all available microphones, or from an appropriate subset of them, are used during the training process in order to obtain the statistical model (GMM) of each class of interest. This is akin to the "multi-style" training procedure, often employed in ASR and other machine learning problems to improve robustness of the produced models. The obtained models can then be used during testing on one or more microphones, in the latter case using the decision fusion framework discussed below.

2.2. Signal fusion

In this approach, a plain delay-and-sum beamformer with no post-filtering is employed to combine audio from multiple microphones into a single enhanced signal (typically, a subset of the available microphones is exploited that are closely located within microphone arrays). For this purpose, the "BeamformIt" software is used [13]. Depending on which channels are combined, one or more beamforming signals can be created, thus also allowing multi-channel training and/or decision fusion approaches to be employed.

2.3. Decision fusion

In this approach, the available class models are tested on the appropriate channels that are to be fused at the decision level. Typically, for example, a single-channel classifier is tested on the respective channel that it is trained on; a multi-channel model is tested on any channel within the set of microphones that is trained on; and a signal-fusion model is tested on its corresponding enhanced signal. Such tests provide sequences of log-likelihood scores for each class and channel of interest, which are then fused at the frame level by one of the methods described next.

2.3.1. Combination strategies

Unweighted log-likelihood sum ("u-sum"): For the current feature frame, the sum of the log-likelihoods over all channels to be fused is computed for each class of interest, thus providing the fused class log-likelihoods for this frame.

Weighted log-likelihood sum ("c-sum"): Similar to the above, but with a confidence-weighted sum of the current frame log-likelihoods over all channels computed for each class instead. The weights are based on channel confidence estimates, calculated as discussed later.

Global log-likelihood maximum ("u-max"): The channel achieving the highest frame log-likelihood over all channels and over all classes is the one chosen to provide all fused class log-likelihoods at the current frame.

Global log-likelihood maximum confidence ("c-max"): The channel with the highest confidence (computed as discussed below) is the one chosen to provide all fused class log-likelihoods at the current frame.

Unweighted majority voting ("u-vote"): At the current frame, and for each channel, the class that ranks first, i.e., achieves the highest frame log-likelihood score over the classes of interest for the particular channel, obtains a vote of one (the other classes obtain a vote of zero). The votes are summed across all channels to be fused, and the class with the highest score (number of votes) is chosen for the current frame.

Weighted majority voting ("c-vote"): As above, but with each vote weighted by its corresponding channel confidence.

2.3.2. Confidence estimation

Approaches "c-sum", "c-max", and "c-vote" require channel confidence estimation to yield necessary weights. Similarly to [14], we utilize, for this purpose, the following channel decision confidence or channel quality indicators.

N-best average log-likelihood difference: For every channel, this is derived by computing the average of the differences in the log-likelihood score between the highest scoring class GMM and the $N-1$ following in descending order (where N is upper bounded by the number of available classes). Large values of this difference indicate high confidence.

N-best average log-likelihood dispersion: This constitutes a modification of the above, where log-likelihood differences between all top N-scoring class pairs are averaged. As before, large values demonstrate high confidence.

Log-likelihood score entropy: The entropy over the probability distribution of all class posteriors is computed. Small entropy values indicate high classification confidence.

Segmental signal-to-noise-ratio (SNR): This is a commonly used channel quality indicator, with high SNR values indicating good data quality.

After experimenting with the above channel confidence indicators, we converged to using "segmental SNR" for AED and the "2-best log-likelihood difference" for VAD, yielding weights after their normalization over the channels fused.

2.4. Feature extraction

Regarding the features used, 13 MFCCs with Δ's and $\Delta\Delta$'s were extracted from the single-channel or fused signal, over 25/100ms duration frames with a 10/20ms shift, for VAD and AED, respectively. In addition, and similarly to [7], we employ as features TDOAs between pairs of adjacent microphones, as these are related to the source location and possibly the class of certain acoustic events. Such features are used to train a separate GMM, which is then combined with MFCC-trained GMMs employing decision fusion.

3. DETECTION APPROACHES

Two detection systems are developed, employing at their core the trained GMMs with multi-channel fusion.

3.1. Viterbi decoding over entire sequence

We denote by $b_{mj}(\mathbf{o}_t)$ the GMM log-likelihood score of event j for the microphone-m at time frame t. In the single-microphone case, using the Viterbi algorithm, the maximum log-probability of observing vectors $\mathbf{o}_1$ to $\mathbf{o}_t$ at microphone m and being in state (event) j at time (frame) t is:

$$\delta_{mj}(t) = \max_i \{\delta_{mi}(t-1) + \log(a_{ij})\} + b_{mj}(\mathbf{o}_t), \qquad (1)$$

where a_{ij} denotes the transition probability from state i to state j. By adding a constant value on the diagonal of the transition matrix , we can tune the flexibility of the decoder to change states (state transition penalty).

To apply our decision fusion approaches using the Viterbi decoding algorithm, we transform the above equation to use the multi-channel log-likelihoods $c_j(\mathbf{o}_t)$ instead of the single-channel $b_{mj}(\mathbf{o}_t)$. These multi-channel log-likelihoods are produced using the fusion methods presented earlier, yielding

$$\delta_j(t) = \max_i \{\delta_i(t-1) + \log(a_{ij})\} + c_j(\mathbf{o}_t). \qquad (2)$$

Majority-voting approaches cannot be used as above. Instead, we can apply the majority-voting scheme at each frame t using the log-likelihood scores $\delta_{mj}(t)$ produced by Viterbi decoding for each microphone m.

3.2. GMM scoring over sliding window

In this approach, detection is performed by sequential classification over sliding windows of fixed duration and overlap. For a given time window and a microphone m, the log-likelihood scores for each event are computed by adding the log-likelihood scores of the individual observations $\mathbf{o}_1, ..., \mathbf{o}_T$ contained in that window: $b_{mj}(\mathbf{o}_1, ..., \mathbf{o}_T) = \sum_{t=1}^{T} b_{mj}(\mathbf{o}_t)$, i.e., the observations are considered independent.

This procedure is performed for every model separately, and then decisions are fused for each sliding window with the methods aforementioned. Concerning the window length and shift, we finally used 0.6/0.4s duration frames with 0.4/0.2s shifts for AED and VAD respectively.

4. EXPERIMENTS AND RESULTS

4.1. AED

Concerning the AED task, development and evaluation of the various approaches is performed on the UPC-TALP multi-microphone corpus of acoustic events [15]. This database contains a set of isolated acoustic events that occur frequently in a meeting environment scenario. In our task, in addition to silence, we have 12 different events in total: knocks (door, table), door slams, steps, chair moving, spoon, paper work, key jingle, keyboard typing, phone ringing, applause, cough and speech. Audio data from a total of 24 channels are available, provided by six T-shaped microphone arrays located on the room walls.

As the UPC-TALP database recordings are divided into 8 independent sessions, experiments have been conducted in a leave-one-out session fashion, keeping seven sessions for training and leaving one for testing.

The results for the AED task are depicted in Table 1. Performance of the various combination schemes considered is reported in terms of DER (Diarization Error Rate) [16], which in our case (isolated events) practically corresponds to frame misclassification. The results presented correspond to the best combination of parametres used (state transition penalty, number of Gaussians). As a baseline in our experiments, the "best estimated-SNR channel" selection strategy (per session) has been considered. For a given session, the SNR for each channel is computed as the ratio between the total energy in the non-silence and silence segments detected. In the "best actual-SNR" method, segment boundaries are given from the ground-truth. In the "oracle best channel" method in each session the channel with the lowest DER is selected. Finally "average over channels" refers to the mean DER of all the single channels results in the leave-one-out experiment.

Concerning the results, at first we observe that Viterbi decoding (HMM) outperforms the sliding window approach (GMM). Regarding the decision-level fusion, we can observe its superiority over the baseline systems.The best approach is "c-sum" which achieves a 8.10% relative error reduction (from 14.20% to 13.05%) compared to the best SNR single-channel system.

The combination of decision fusion with multi-channel training and signal fusion, yielded no improvement. Yet, the results remained better than the single-channel baseline. Finally, the combination of TDOAs with MFCCs GMMs in the decision level obtained the best overall result (Table 2). In particular, the combination of TDOAs with the "u-sum" method yielded a 12.88% DER, which corresponds to a 9.30% relative error reduction from the "best estimated-SNR channel" approach (Fig. 1), and 11.20% from the "average over channels" DER. This can be explained by the fact that some events occur in similar locations in the various sessions. The best combination gave weight equal to 0.1 to TDOAs

training style		single-channel		multi-channel	signal fusion
trained models (#)		24		1	6
channels tested (#)		24		24	6
model type		GMM	HMM	HMM	HMM
best estimated-SNR channel		18.54	14.20	14.59	14.53
best actual-SNR channel		18.43	14.16	14.43	14.48
average over channels		19.21	14.34	14.42	14.42
oracle best channel		17.50	12.71	13.04	13.40
decision fusion	u-max	18.94	13.76	14.19	14.37
	u-vote	18.09	13.13	13.42	13.40
	u-sum	17.91	13.21	13.36	13.50
	c-max	18.21	13.66	13.96	14.15
	c-vote	18.12	13.17	13.50	13.78
	c-sum	17.94	13.05	13.29	13.43

Table 1. Multi-channel fusion results for the AED problem. Results are depicted in DER %.

TDOAs & MFCCs		AED	VAD
decision fusion	u-sum	**12.88**	**3.90**
	c-sum	12.92	4.00

Table 2. Results for the fusion of MFCC and TDOAs models for AED and VAD tasks.

model and 0.9 to MFCCs model. Regarding the DER of TDOAs model (without fusion) it reaches 36.64%.

In order to verify that the improvement observed by the multi-channel approaches is statistically significant, we apply the Wilcoxon signed-rank test. In particular, a one-sided Wilcoxon test [17] is performed to compare the detection accuracies over all 8 leave-one-out experiments between the various multi-channel approaches and the baseline system. We also compare the significance of improvement between weighted and non-weighted approaches.

The outcomes of the tests are positive using the value $p < 0.05$. The improvements over the baseline observed are judged as significant in most approaches (“TDOAs” , “c-sum”, “u-sum”, “c-vote”, “u-vote”, “c-max”) (all except for the “u-max” method). Also statistical significant improvement was observed between “c-sum” and “u-sum” methods. This indicates that the weighted approach performs slightly but steadily better than the simple one.

4.2. VAD

We perform our VAD experiments in the DIRHA simulated corpus [18] designed for the purposes of DIRHA project [19] at FBK. This database contains speech commands and dialogs occurring in an apartment comprising 5 different rooms. A big variety of acoustic events also can happen in different locations of the apartment and often overlap with speech. The audio data contains simulated recordings from 40 microphones placed in the walls and ceilings of the rooms. In total, 150 simulations of 1 minute duration each were generated by convolving pre-recorded data with the impulse response of the apartment for different locations. In our experiments we use 75 simulations for training and the rest for testing.

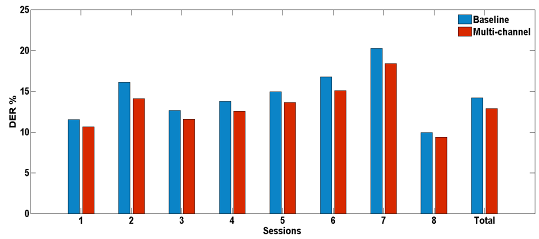

Fig. 1. Performance (in DER %) of baseline “best estimated-SNR” and best multi-channel approach (“TDOAs & MFCCs”) for the 8 sessions of AED problem.

In the VAD task, we have experimented with the same fusion schemes as in AED. Multi-channel training was performed on each of the 5 rooms of the apartment. From the results in decision fusion (Table 3), we can immediately observe the superiority of multi-channel approaches vs. the single-channel case. Also, in this task, multi-channel training helped increasing the performance of the system. Finally in the combined system, non-negligible improvements are observed by using the weighted approaches. This is not surprising, since the employed channel confidence metric seems to provide a good indication of its decision correctness, as depicted in Fig. 2.

The best result is obtained again with the combination of TDOAs and MFCCs GMM models in the decision level. It achieves a 44.84% relative detection error reduction (from 7.07% to 3.90%) compared to the “best estimated-SNR channel” and 52.55% from the “average over channels” method (from 8.22% to 3.90%). Regarding the DER of TDOAs model alone it reaches 23.02%.

The VAD task, has lower complexity than AED, as it considers only 2 classes. Although, in our experiments, the environment of VAD problem was much more challenging than that of AED, as it is comprises 5 rooms, and contains various background noise sources overlapping with speech and located in different positions of the apartment. This kind of the environment revealed more the utility of multi-channel fusion approaches.

In correspondence to the AED problem, we tested the significance of the multi-channel approaches over the baseline. Similarly with AED, the result was positive for all multi-channel methods except “u-max” method. However, the improvements of weighted over non-weighted methods were not judged as significant ones.

5. CONCLUSIONS

In this paper, we investigated multi-channel combination approaches in different levels for the problems of acoustic event and voice activity detection. In both problems, and especially in VAD, multi-channel approaches outperformed the baseline single-channel system.

training style		single-channel		multi-channel	signal fusion
trained models (#)		40		5	14
channels tested (#)		40		40	14
model type		GMM	HMM	HMM	HMM
best estimated-SNR channel		7.93	7.07	5.45	4.85
best actual-SNR channel		5.32	4.08	4.30	4.04
average over channels		9.41	8.22	8.01	7.87
oracle best channel		3.52	1.67	1.57	2.11
decision fusion	u-max	11.01	9.74	10.24	10.06
	u-vote	6.72	5.50	5.44	5.39
	u-sum	6.24	4.28	4.38	4.55
	c-max	6.28	5.10	4.70	4.34
	c-vote	5.97	4.97	4.84	4.83
	c-sum	6.06	4.41	4.18	4.29

Table 3. Multi-microphone fusion results for VAD problem. Results are depicted in DER %.

Concerning the back-ends used, we can observe that Viterbi decoding is more appropriate for the detection task. It finds the most probable sequence of events in an optimal way. As for the decision fusion approaches, in general summation methods work better than majority based ones, and weighted better than unweighted ones. Finally, the extraction of the TDOAs and the training of a separate GMM model with them increased further the performance of the overall system.

Finally we must underline that the usefulness and contribution of multi-channel approaches is better demonstrated in larger environments and under more adverse conditions.

In future work, we will investigate the more general problem of overlapped acoustic event detection in noisy smart home environments. We will also experiment with better confidence metrics in order to improve further the performance of the weighted decision fusion approaches.

Acknowledgments

The authors would like to thank Prof. C. Nadeu and the TALP group at Universitat Politecnica de Catalunya (UPC) for the free distribution of the UPC-TALP database. We would also like to acknowledge Dr. M. Omologo and the SHINE group at FBK for providing us with the simulated DIRHA corpus.

REFERENCES

[1] C. Zieger, "An HMM based system for acoustic event detection," in [16], pp. 338–344. Springer, 2008.

[2] M. Baillie and J.M. Jose, "Audio-based event detection for sports video," in *Image and Video Retrieval*, pp. 300–309. Springer, 2003.

[3] X. Zhuang, X. Zhou, A. Hasegawa-Johnson, and T. S. Huang, "Real-world acoustic event detection," *Pattern Recognition Let.*, 31(12):1543–1551, 2010.

[4] T. Butko, F. G. Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: Online implementation in a smart-room," in *Proc. EUSIPCO*, 2011.

[5] A. Temko, C. Nadeu, and J.I. Biel, "Acoustic event detection: SVM-based system and evaluation setup in CLEAR'07," in [16], pp. 354–363. Springer, 2008.

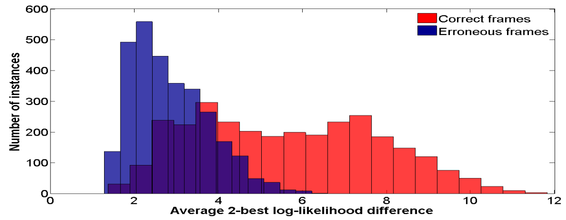

Fig. 2. Histograms of average value of confidence (2-best log-likelihood difference) for both correctly and incorrectly classified frames in the VAD problem. For each simulation and for each microphone we compute one value for mean confidence of erroneous frames and one for correct frames.

[6] T. Heittola and A. Klapuri, "TUT acoustic event detection system 2007," in [16], pp. 364–370. Springer, 2008.

[7] T. Butko, A. Temko, C. Nadeu, and C. C. Ferrer, "Fusion of audio and video modalities for detection of acoustic events," in *Proc. INTERSPEECH*, 2008, pp. 123–126.

[8] Q. Li, J. Zheng, A. Tsai, and Q. Zhou "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. on Speech and Audio Process.*, 10(3):146–157, 2002.

[9] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using MFCC features and Support Vector Machine," in *Proc. SPECOM*, 2007, vol. 2, pp. 556–561.

[10] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program.," in *Proc. INTERSPEECH*, 2012.

[11] E. Marcheret, G. Potamianos, K. Visweswariah, and J. Huang, "The IBM RT06s evaluation system for speech activity detection in CHIL seminars," in *Machine Learning for Multimodal Interaction*, pp. 323–335. Springer, 2006.

[12] J. E. Rubio, K. Ishizuka, H. Sawada, S. Araki, T. Nakatani, and M. Fujimoto, "Two-microphone voice activity detection based on the homogeneity of the direction of arrival estimates," in *Proc. ICASSP*, 2007, vol. 4, pp. 385–388.

[13] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. on Audio, Speech, and Language Process.*, 15(7):2011–2022, 2007.

[14] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition.," in *Proc. INTERSPEECH*, 2000, pp. 746–749.

[15] T. Butko, C. Canton-Ferrer, C. Segura, X. Giro, C. Nadeu, J. Hernando and J.R. Casas, "Acoustic event detection based on feature-level fusion of audio and video modalities," in *EURASIP, Journal on Advances in Signal Process.*, 2011.

[16] R. Stiefelhagen, R. Bowers, and J. Fiscus, *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007*, vol. 4625, Springer, 2008.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[18] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. LREC*, 2014.

[19] "DIRHA: Distant-speech interaction for robust home applications," [Online] Available at: `http://dirha.fbk.eu/`.