SOUND EVENT DETECTION USING NON-NEGATIVE DICTIONARIES LEARNED FROM ANNOTATED OVERLAPPING EVENTS

Onur Dikmen

Department of Information and Computer Science,
Aalto University, 00076, Finland
onur.dikmen@aalto.fi

Annamaria Mesaros

Department of Signal Processing and Acoustics,
Aalto University, 00076, Finland
annamaria.mesaros@aalto.fi

ABSTRACT

Detection of overlapping sound events generally requires training class models either from separate data for each class or by making assumptions about the dominating events in the mixed signals. Methods based on sound source separation are currently used in this task, but involve the problem of assigning separated components to sources. In this paper, we propose a method which bypasses the need to build separate sound models. Instead, non-negative dictionaries for the sound content and their annotations are learned in a coupled sense. In the testing stage, time activations of the sound dictionary columns are estimated and used to reconstruct annotations using the annotation dictionary. The method requires no separate training data for classes and in general very promising results are obtained using only a small amount of data.

Index Terms— Non-negative matrix factorization, Sound event detection

1. INTRODUCTION

Sound event detection is an important part of Computational Auditory Scene Analysis (CASA). The specific task is content analysis of audio – extracting information from an audio recording, to represent its content in terms of sound events taking place. This can be tailored to specific applications in healthcare, such as monitoring of patients or elderly people [1]; security surveillance – detecting gun shots [2] or other type of violence related events; wildlife monitoring in ecology [3], and others. The target audio data for such applications is our everyday environment, which is a complex mixture of sound sources representing the different sound events. Detection, identification and segregation of the sound events in such mixtures is a natural thing for humans, but a challenging problem for automatic algorithms.

Early work considered the mixture signal as a sequence of events, and detected only the prominent sound event at each time, to output a single sequence of events. Methods for prominent event detection are based on supervised classification, and train class models for the sound events involved in the classification. In artificial cases, examples of sound events may be available in isolation, and in this case training models for each class is straightforward. For training models from the mixture signal, authors of [4] considered the segments of the mixture audio annotated to each class, under the assumption that the nonrelevant parts will average out during training. Another approach for training from mixture signals was presented in [5]: the authors consider regions with overlapping events as a separate class and train models for individual sound event classes and multiple combinations of them. Such a system can easily become untrainable, when there are too many combinations involved.

Recent approaches considered detection of multiple overlapping events in the mixtures. For example, in [6], event class models were estimated using corresponding segments of mixture audio as recorded from the natural environment, and the multiple overlapping events were obtained by successively applying Viterbi algorithm to decode the next-best sequence of events.

For dealing with the overlapping events at the model training stage, sound source separation methods provide a promising solution. When the number of sources is known and assumptions about them can be made reliably, the resulting separated audio components correspond to the involved sound sources [7], and the representative audio content can be used in training a model for a specific class. However, without prior knowledge, it is not easy to assign the separated components to sources. Methods for choosing a representative stream were proposed in [8], based on an expectation-maximization type algorithm for iterative selection within the streams resulted from non-negative matrix factorization. The mentioned studies were built on the classical supervised classification approach, and tackled the acoustic model construction from mixture signals by assigning audio content to classes. This approach cannot completely overcome the problem of training models from mixture signals.

We propose a novel approach for sound event detection, based on a sparse dictionary representation of the mixture signal. The MMLE algorithm is used to learn non-negative dictionaries [9] for representing the audio data based on the spectrogram of the mixture signal and its annotation represented in a frame-level activity matrix form. The obtained audio dictionary is then used to estimate the annotation matrix of a test signal. The method is used without assigning the dictionary components to sound sources, and is therefore unsupervised in the sense of not building any specific models for event classes. A major advantage of the proposed method is that it works directly on the mixture signal, and provides directly a polyphonic annotation matrix, so there is no need of obtaining isolated examples of sound events, or separating audio into classes, as in the earlier approaches. Another advantage of this method is the possibility of learning components from a small amount of data.

The paper is organized as follows: Section 2 presents the non-negative matrix factorization framework and the details of the proposed method for estimating the annotation matrix of test data. Section 3 describes the experimental setup for evaluation, the database and the metrics. Section 4 presents the experimental results and discussion, and finally Section 5 presents conclusions and future work.

This work is supported by The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170).

2. METHOD

The proposed method is based on the idea that spectrogram and corresponding annotation matrix share information about activation of sound sources in time. Non-negative matrix factorization (NMF) [10], already proven very effective in many audio processing tasks, e.g., source separation, transcription, etc., is now a natural tool to extract audio templates (dictionary) and their excitations in time. This work assumes a single excitation matrix for the audio spectrogram and the annotation matrix and learns two dictionaries corresponding to these modalities, i.e., data. When a test recording arrives, its excitation matrix corresponding to the audio dictionary is estimated and annotations are estimated by multiplying the annotation dictionary with this excitation matrix.

The objective in NMF is to find a low-rank approximation to the observed data by expressing it as a product of two non-negative matrices, i.e., $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{WH}$ with $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$. This objective is pursued through the minimization of an information divergence between the data and the approximation, i.e., $D(\mathbf{V}||\hat{\mathbf{V}})$. The divergence can be any appropriate one for the data/application such as Kullback-Leibler (KL), Itakura-Saito (IS), Euclidean distance or families of divergences like β (for which, all three divergences above are special cases), α, γ, Rényi, etc. There are very efficient algorithms for these divergences based on multiplicative update rules [10, 11, 12].

The augmented model considered here (Fig. 1a) consists of the spectrogram $\mathbf{V}^{(1)}$ of size $F \times N$ and the annotation matrix $\mathbf{V}^{(2)}$ of size $E \times N$, and the unknown dictionaries $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, and the unknown common excitation matrix, $\mathbf{H}$. Here, F is the number of frequency bins, N is the number of frames, E is the number of events. The objective function to be minimized is $\eta_1 D^{(1)}(\mathbf{V}^{(1)}||\mathbf{W}^{(1)}\mathbf{H}) + \eta_2 D^{(2)}(\mathbf{V}^{(2)}||\mathbf{W}^{(2)}\mathbf{H})$, i.e., the divergences used can be different and weighted (with η_i). This coupled matrix factorization problem can be effectively solved for β-divergence using generalized linear models [13]. In this paper, we handle the case by considering both $D^{(1)}$ and $D^{(2)}$ to be KL and $\eta_1 = \eta_2 = 1$ and use an estimator based on maximum marginal likelihood (MMLE) [9], which was shown to return sparse solutions and avoid overfitting.

In MMLE for KL divergence, the expansion coefficients h_{kn} are assumed to be random variables with Gamma prior, such that $h_{kn} \sim \mathcal{G}(h_{kn}|\alpha_k, \beta_k)$, where $\mathcal{G}(x|\alpha,\beta) = \beta^\alpha/\Gamma(\alpha)\, x^{\alpha-1} \exp(-\beta\, x)$, $x \geq 0$. No constraint is imposed on $\mathbf{W}$, i.e., it remains a free deterministic parameter. By deriving a variational expectation-maximization (EM) algorithm on this model, the following multiplicative update rules for w_{fk} are obtained

$$w_{fk} \leftarrow w_{fk} \frac{\sum_n \exp(\langle \log h_{kn}\rangle) v_{fn}/[\mathbf{W}\exp(\langle \log \mathbf{H}\rangle)]_{fn}}{\sum_n \langle h_{kn}\rangle},$$

where $\langle . \rangle$ denotes expectation w.r.t the variational distribution.

The experiments in [9] showed that MMLE indeed is not prone to overfitting and optimizes the number of components by assigning redundant dictionary columns to zero. Thus, choosing a large value for K is enough to obtain dictionaries without noise templates or duplicates.

The goal in the training step is to learn the dictionaries $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$. The learned excitation matrix $\mathbf{H}_{\text{train}}$ is simply discarded because it only contains information about the excitation of the dictionaries on the training data. Testing involves learning the excitation matrix $\mathbf{H}_{\text{test}}$ for the test recording using only the spectrogram data (Fig. 1b). Then, the estimation of the annotation matrix is done by thresholding $\mathbf{W}^{(2)}\mathbf{H}_{\text{test}}$. The overall algorithm is given in Algorithm 1.

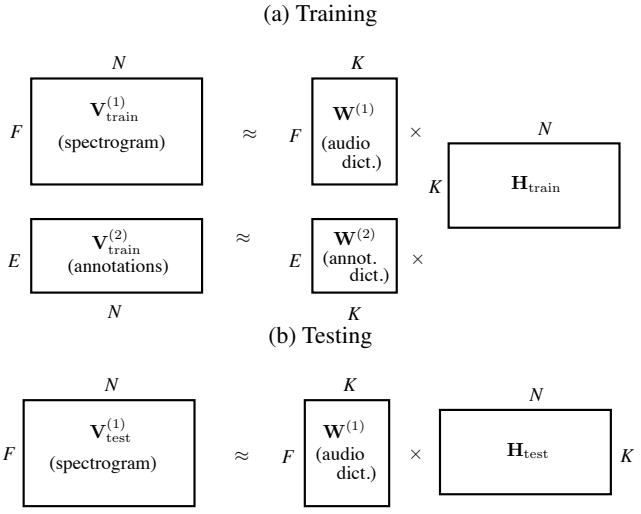

Figure 1: Models used for training and testing.

Algorithm 1 Pseudocode for the method.

for each training pair $\mathbf{V}_i^{(1)}$ and $\mathbf{V}_i^{(2)}$ **do**
  Estimate $\mathbf{W}_i^{(1)}$, $\mathbf{W}_i^{(2)}$, and $\mathbf{H}_i$
  Discard $\mathbf{H}_i$
end for
$\mathbf{W}^{(1)} = [\mathbf{W}_1^{(1)}\mathbf{W}_2^{(1)}\cdots\mathbf{W}_I^{(1)}]$
$\mathbf{W}^{(2)} = [\mathbf{W}_1^{(2)}\mathbf{W}_2^{(2)}\cdots\mathbf{W}_I^{(2)}]$
for each test recording $\mathbf{V}_j^{(1)}$ **do**
  Estimate $\mathbf{H}_j$ with fixed $\mathbf{W}^{(1)}$
  Calculate $\hat{\mathbf{V}}_j^{(2)} = \mathbf{W}^{(2)}\mathbf{H}_j$
  Threshold $\hat{\mathbf{V}}_j^{(2)}$ to obtain annotations
end for

3. EXPERIMENTAL SETUP

The proposed method will be evaluated using the audio recordings provided in the AASP challenge Detection and Classification of Acoustic Scenes and Events [14]. The challenge provided an evaluation framework on the tasks of modeling and identifying acoustic scenes containing non-speech and non-music and detecting audio events. We perform one set of experiments as a leave-one-out test, with one recording being kept as test data and all the others being used in learning the non-negative dictionaries. To demonstrate the capabilities of the method in learning the dictionaries from a small amount of data, we perform a second experiment by learning the dictionaries from a single recording and testing all the other ones.

3.1. Audio Events Database

The training dataset provided for the Event Detection - Office Synthetic task of the AASP challenge consisted of recordings of indi-

vidual events. Our method is designed to estimate the dictionary directly from overlapping events data, therefore this training dataset was not suitable. The development set provided for the task consists of nine recordings built by sequencing recordings of individual events and background recordings.

Event classes present in the data are: door knock, door slam, speech, human laughter, clearing throat, coughing, drawer, printer, keyboard click, mouse click, object (specifically pen, pencil or marker) put on table surfaces, switch, keys (put on table), phone ringing, short alert (beep) sound, page turning. The synthetic scenes were generated by randomly selecting for each event occurrence one representative excerpt from natural scenes, and mixing all those samples over a background noise. The scenes were mixed at different density of events (low, medium, high). The average SNR of events over the background noise is also specified (6dB, 0dB,-6dB), and the same level is used for all events. The synthesized scenes are mixed down to mono and provided with accompanying ground-truth annotations. The set consists of nine recordings (9 combinations, 3 SNR cases x 3 event density cases), each of length 1:30 min, containing between 6 and 73 events.

3.2. Performance evaluation metrics

Evaluation is done using the metrics provided in the same AASP challenge, in order to set comparison grounds for the method. The metrics include a frame-based, event-based and class-wise F-score, with event-based and class-wise taking into account onset only or both onset and offset of events. The frame based evaluation is performed in 10 ms steps, by calculating precision and recall in 10 ms frames, and reporting the balanced F-score. Additionally, frame based acoustic event error rate is defined with respect to deletions (D), insertions (I) and substitutions (S) for the ground truth number of events (N) in each 10 ms frame:

$$AEER = (D + I + S)/N \cdot 100$$

Event-based evaluation considers F-scores averaged over event instances, in two metrics. Onset-only evaluation considers an event correctly detected if its onset is within 100 ms tolerance window. Onset-offset evaluation considers and event correctly detected if its onset is within a 100ms tolerance window and its offset is within 50% range of the ground truth events offset with respect to the duration of the event. Class-wise evaluation is done by calculating metrics separately for each class and averaging over a recording.

4. RESULTS

The estimated annotation matrix of the test recording is obtained by calculating the product $\hat{\mathbf{V}}_j^{(2)} = \mathbf{W}^{(2)}\mathbf{H}$, where $\mathbf{H}$ is the minimizer of $D_{\mathrm{KL}}(\mathbf{V}^{(1)}||\mathbf{W}^{(1)}\mathbf{H})$, and $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are the dictionary matrices learned and fixed at the training step. Postprocessing of the result includes class-wise normalization (each row) and binarization by assigning elements greater than $1/N$ to one and the rest to zero. Then, only continuous blocks of at least 20 frames are considered in order to obtain continuous activity patterns for each event class. The choice of the value 20 is arbitrary, a subjective lower bound on the duration of single events.

Figure 2 presents the ground truth and the estimated annotations for one test case with medium density of events. Average F-scores of the metrics described in Section 3 are given in Table 1 for the three SNR levels and the overall average. The numbers in each SNR column represent the average of F-scores over three test cases corresponding to the given SNR value. The frame-based F-score is consistently good in all the SNR situations. The frame-based performance can be compared with the block-wise metric presented in [6], only the block size is 100ms. As can be noticed from [6], the smaller the block, the lower the performance, due to events onset and offset being at incorrect time positions. It follows that our method surpasses greatly their performance, by obtaining over 50% F-score on very small blocks, compared to 20% (one second blocks) and 30% (30 second blocks). It is hard to compare results obtained on datasets, however in both cases the complexity of data in number of classes is similar (16).

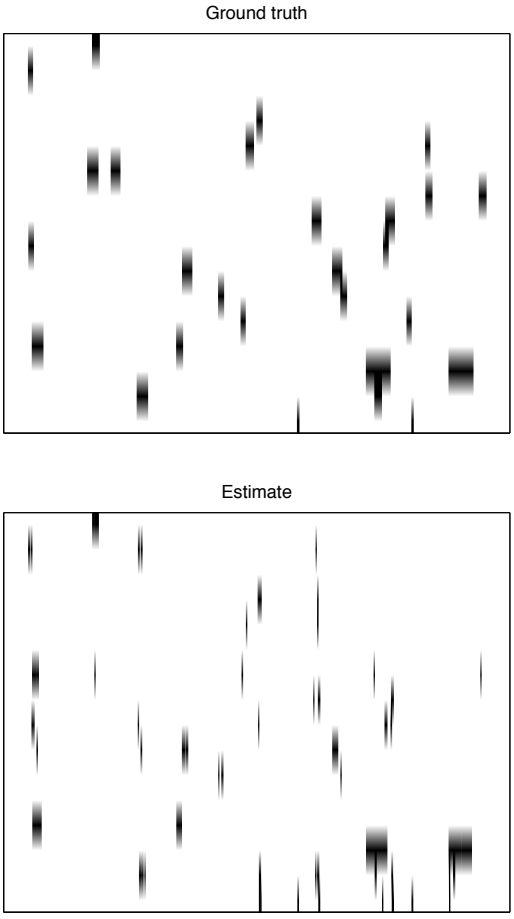

Figure 2: Ground truth (top) and estimated (bottom) annotations for SNR 0 case with medium density of events.

The method offers the possibility of learning dictionaries from a small amount of annotated audio, as long as it contains all the event classes expected in the test data. We chose the high events density file with SNR 0 dB as a representative middle ground for the learning stage. The average F-scores for estimated annotation of the other eight files are presented in Table 2, separately for SNR levels and the overall average. In this case the results in the SNR0 column are the average of only two files (one low and one medium events density). Frame-based average F-score decreases by 7 percentage units. In this situation the dictionary is learned from only 90

Table 1: Leave-one-out F-scores

	SNR 6	SNR 0	SNR -6	Average
Frame-based	0.579	0.578	0.481	0.546
Event-based (on)	0.257	0.290	0.186	0.244
Event-based (on-off)	0.154	0.140	0.075	0.123
Class-wise (on)	0.324	0.306	0.231	0.287
Class-wise (on-off)	0.215	0.151	0.095	0.153

Table 2: F-scores with single recording as the training set

	SNR 6	SNR 0	SNR -6	Average
Frame-based	0.499	0.492	0.442	0.478
Event-based (on)	0.213	0.166	0.181	0.187
Event-based (on-off)	0.120	0.069	0.051	0.080
Class-wise (on)	0.232	0.208	0.197	0.213
Class-wise (on-off)	0.149	0.111	0.073	0.111

Table 3: Leave-one-out AEERs

	SNR 6	SNR 0	SNR -6	Average
Frame-based	1.095	0.978	1.043	1.039
Event-based (on)	3.422	2.886	3.087	3.132
Class-wise (on)	3.123	2.624	2.661	2.803

Table 4: AEERs with single recording as the training set

	SNR 6	SNR 0	SNR -6	Average
Frame-based	1.191	1.098	0.997	1.095
Event-based (on)	3.061	3.382	2.756	3.066
Class-wise (on)	2.609	3.033	2.243	2.629

seconds of audio, eight times less than in the previously presented experiment. The results are consistent with the previous ones and follow a similar pattern for the different SNR situations, with the -6 dB recordings having the lowest frame-based performance.

The frame-based acoustic event error rate for the two experiments is presented in Tables 3 and 4. The average values are very similar, which implies the method robustness to the amount of training data.

5. CONCLUSIONS

We presented a method for learning non-negative dictionaries from overlapping annotated sound events and corresponding audio. The dictionaries are meant to be used in sound event detection, with the possibility of obtaining directly an estimation of annotation with overlapping events, just like the material used for learning. This method has the advantage of avoiding the ambiguity in typical unsupervised sound source separation approaches when assigning the separated audio content to the modeled sources. The dictionary learned from the mixture signal and its event based annotation is used as such in an inverse operation, to generate an estimate of events activity without any correspondence between dictionary components and specific events. This fits very well also in the situation of not having separately recorded sound events, as some may simply be impossible to record in isolation. Even with small amount of available annotated data, the method performs quite well, and the general results we obtained are encouraging. An immediate future work would be combining different divergences, e.g., IS for the spectrogram and KL for the annotations, in the objective function. Since IS divergence is better suited for spectrogram modeling, this may be profitable.

6. REFERENCES

[1] Y.-T. Peng, C.-Y. Lin, M.-T. Sun, and K.-C. Tsai, "Healthcare audio event classification using Hidden Markov Models and Hierarchical Hidden Markov Models," in *IEEE International Conference on Multimedia and Expo, 2009 (ICME 2009)*, 2009, pp. 1218–1221.

[2] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *IEEE International Conference on Multimedia and Expo, 2005 (ICME 2005)*, 2005, pp. 1306–1309.

[3] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal of Applied Signal Processing*, vol. 2007, no. 1, pp. 64–64, 2007.

[4] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real-life recordings," in *18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, 2010, pp. 1267–1271.

[5] 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 *19th European Signal Processing Conference (EUSIPCO 2011)*, 2011, pp. 1307–1311.

[6] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech and Music Processing*, 2013.

[7] S. Innami and H. Kasai, "NMF-based environmental sound source separation using time-variant gain features," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1333 – 1342, 2012.

[8] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013.

[9] O. Dikmen and C. Févotte, "Maximum marginal likelihood estimation for nonnegative dictionary learning," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP'11)*, Prague, Czech Republic, 2011.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[11] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorization*. John Wiley and Sons, 2009.

[12] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, 2011.

[13] K. Y. Yilmaz, A. T. Cemgil, and U. Simsekli, "Generalized coupled tensor factorization," in *NIPS*, 2011.

[14] "IEEE AASP Challenge: Detection and classification of acoustic scenes and events." [Online]. Available: http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/