SOUND EVENT DETECTION IN REAL LIFE RECORDINGS USING COUPLED MATRIX FACTORIZATION OF SPECTRAL REPRESENTATIONS AND CLASS ACTIVITY ANNOTATIONS

Annamaria Mesaros[1], *Toni Heittola*[1], *Onur Dikmen*[2], *Tuomas Virtanen*[1]

[1] Department of Signal Processing, Tampere University of Technology
[2] Department of Information and Computer Science, Aalto University

ABSTRACT

Methods for detection of overlapping sound events in audio involve matrix factorization approaches, often assigning separated components to event classes. We present a method that bypasses the supervised construction of class models. The method learns the components as a non-negative dictionary in a coupled matrix factorization problem, where the spectral representation and the class activity annotation of the audio signal share the activation matrix. In testing, the dictionaries are used to estimate directly the class activations. For dealing with large amount of training data, two methods are proposed for reducing the size of the dictionary. The methods were tested on a database of real life recordings, and outperformed previous approaches by over 10%.

Index Terms— coupled non-negative matrix factorization, non-negative dictionaries, sound event detection

1. INTRODUCTION

Sound event detection has an important position in Computational Auditory Scene Analysis [1], with a specific purpose of recognizing and locating individual sounds, generally referred to as sound events. Recognizing sound events in audio has a large applicability, for healthcare and general automatic surveillance [2, 3], and other general purposes, such as detecting interesting segments in videos based on audio [4, 5].

Simplified situations for sound event detection include classification of isolated sounds [6], and detection of isolated sounds presented as a sequence [7]. This is not realistic for many applications, considering the complexity of the audio surrounding us in everyday life, but detection of overlapping sounds continues to be a challenging problem.

Sound event detection in complex mixtures started by considering the signal as a sequence of most prominent sounds, to detect a single sequence of events. The task was often solved using supervised classification approaches, with event class models trained from isolated data [8] or directly from the mixtures [9, 10, 11]. Two different approaches of the model training method are observed: in [10], annotated segments were aggregated for each event class, with the assumption that overlapping sound events contribution to these segments will average out during training, and the resulting model will be characteristic mainly to the class under training; in [9, 11], regions with overlapping events were considered as separate class and detected accordingly.

More recent approaches considered detection of overlapping sound events, bringing the applicability of methods closer to real life situations. Using event classes trained from mixture audio as in [10], multiple sequences of events were obtained through successive application of Viterbi algorithm [12]. More promising approaches include sound source separation methods as preprocessing stage in the model training phase. However, there is a difficulty in assigning the separated components to the sound sources. This requires making assumptions about the number of sources and their dynamics [13], or using classification methods to establish the sound source prominence and select the most representative component out of the separated ones [14].

In our previous work, we proposed an approach for overlapping sound event detection that overcomes the need to assign separated components to sound event classes [15]. The method is based on learning non-negative dictionaries through joint use of spectrum and class activity annotation of the training data. The learned dictionaries are then used to estimate the class activity annotation for each event class for a test signal. The advantage of this method is its application directly on the mixture signal, and output presented directly in the form of an annotation containing overlapping events. As such, this approach completely avoids the need for isolated examples of sound events, or creating models for event classes in a supervised way, as used in earlier approaches. The method was tested on a small synthetic dataset, with very encouraging results among available methods [16].

In the present paper, we extend this method to deal with realistic data, meaning a large amount of real life recordings, with high complexity of audio content in number of sources and overlapping sounds. For long recordings, directly applying the proposed method is computationally infeasible, as the resulting dictionary grows with the amount of data. We propose two different methods to handle the amount of data by reducing the size of the non-negative dictionary obtained in training: reducing the number of dictionary components, and reducing the length of the individual components.

The paper is organized as follows: Section 2 presents the details of the training procedure and the methods for reducing the size of the dictionary. Section 3 describes the experimental setup for evaluation and the results, Section 4 presents the discussion and comparison of results to previous work. Section 5 presents conclusions and future work.

2. METHOD

The event detection method is based on Non-negative Matrix Factorization (NMF). Non-negative dictionaries are obtained from all available training data, and are then used to estimate the class activity matrix for the test data, that is interpreted as annotation.

2.1. Non-negative dictionaries

The method is based on coupled NMF [17], where two different data sets share some factors, i.e., activation in time. In our problem, the two matrices are the audio frequency spectrum and a matrix of frame-level binary class presence indicators obtained from the annotation of the audio data. NMF has long been a standard dictionary learning tool in many audio processing tasks, such as source separation, transcription, etc.

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., $\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\mathbf{H}$ 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 [17, 18, 19].

The coupled NMF problem considered is to minimize

$$\eta_1 D^{(1)}(\mathbf{V}^{(1)}||\mathbf{W}^{(1)}\mathbf{H}) + \eta_2 D^{(2)}(\mathbf{V}^{(2)}||\mathbf{W}^{(2)}\mathbf{H})$$

w.r.t. dictionaries $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, and common excitation matrix, $\mathbf{H}$; where $\mathbf{V}^{(1)}$ is the spectrum of size $F \times N$ and $\mathbf{V}^{(2)}$ is the class activity annotation matrix of size $E \times N$. Here, F is the number of frequency bins, N is the number of frames, E is the number of event classes. η_i are the weights associated with divergences. This coupled matrix factorization problem can be effectively solved for β-divergence using generalized linear models [20]. In this paper, we are interested in sparsest possible dictionaries, so we make use of an estimator based on maximum marginal likelihood (MMLE) [21], which was shown to return sparse solutions and avoid overfitting. This corresponds to choosing $D^{(1)}$ and $D^{(2)}$ to be KL with equal weights, i.e., $\eta_1 = \eta_2 = 1$.

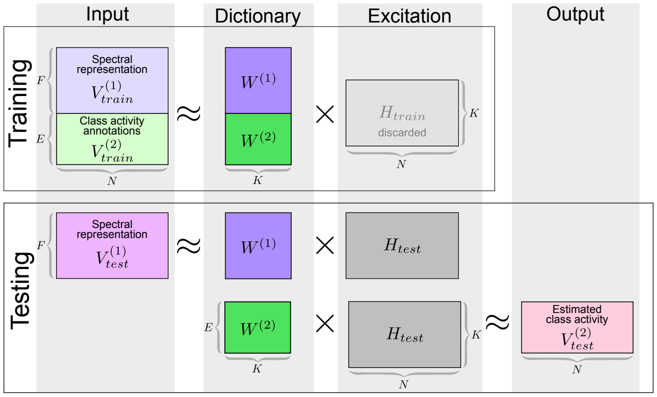

Fig. 1. Overview of training and testing procedure

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 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 its frequency spectrum. Then, the estimated annotation matrix is obtained by calculating the product $\hat{\mathbf{V}}^{(2)} = \mathbf{W}^{(2)}\mathbf{H}_{\text{test}}$, where $\mathbf{H}_{\text{test}}$ is the minimizer of $D_{\text{KL}}(\mathbf{V}^{(1)}||\mathbf{W}^{(1)}\mathbf{H})$, and $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are the dictionary matrices learned and fixed at the training step. The training and testing procedures are illustrated in Figure 1.

2.2. Dictionary size reduction

A straightforward way to deal with long audio recordings is to segment the training and testing data into more manageable length segments, and apply the proposed method by considering each of these segments separately. However, for testing, $\mathbf{W}^{(1)}$ is formed by concatenating all components estimated in training, resulting in a complete dictionary of a significant size. We reduce the dimensionality of the problem by methods that affect the size of this dictionary.

The first method to reduce the size of the dictionary is clustering of the components. As there are multiple segments in training that belong to the same recording, it is likely that their content, and therefore the obtained dictionary, is similar. It follows that we can cluster the complete dictionary and use the centroids of the clusters as new components. This approach has the advantage of offering clear control of the number of components, but has the disadvantage that the clustering has to be repeated when adding new training data.

The second method is using a smaller-dimensional representation of the spectrum, instead of the full-size spectrogram. For this, we chose the mel scale, and transformed the spectrogram into mel magnitudes. Using mel-magnitude and full dictionary allows control of the size of each dictionary component, and keeps the possibility of adding training material by simply concatenating the corresponding components to the existing dictionary.

3. EXPERIMENTS

We present comparative results for three methods: decomposition of full spectrum representation and testing with complete dictionary, decomposition of full spectrum representation and testing with clustered dictionary, and decomposition of mel energy spectrum representation and testing with complete dictionary.

3.1. Experimental setup

The data used to test the performance of the methods is a database of real-life recordings from 10 different contexts: basketball, beach, bus, car, hallway, office, restaurant, shop, street and track&field [10]. Each context was treated separately. This means that the size of the annotation matrix is different for each context, depending on the number of event classes present (9-16).

Training and testing were done using one minute segments that were factorized independently. The spectrum of each one-minute segment of audio was calculated in 100 ms frames with a 50% overlap, using 1024 fft-bins. In training, each one-minute segment contributes to the complete dictionary by a number of components automatically selected by the algorithm. For testing, the class activity matrix of each one minute segment was estimated separately, then the ones corresponding to a single recording were concatenated in their corresponding order, to provide the full class activity estimation. This estimated class activity matrix was then evaluated against the ground truth annotation. Training and testing were done in five folds, respecting the train/test partitioning used in previous experiments on the same data. One fold was used as development set for determining the best number of clusters, but for comparison purposes the final results are presented as average performance of all five folds.

Evaluation was done using the metric proposed in [12], with a block length of one second, to allow an exact comparison with performance of previously developed systems. The metric provides a block-level detection accuracy rather than exact onset-offset detection information. Within a one-second length block, an event is regarded as correctly detected if it has been detected somewhere within the block and the same event label also appears in the annotations within the same block. The block-wise detection accuracy is represented as the balanced F-score between precision and recall calculated in each block. For each recording, average of block-wise F-score was calculated.

The estimated class activity matrix obtained after the two-step testing procedure is not binary, and it needs to be processed in order to be interpreted as presence/absence of sounds. We consider the numbers in this estimated class activity annotation matrix as representing the "amplitude" of each event class in the corresponding timeframe, and we threshold them by the mean value of the matrix: values under the mean are considered 0, values over the mean are considered 1. Furthermore, we keep only sequences of at least 200 ms duration, discarding all presence indicators that would result in sound events shorter than that.

dictionary size	full	100	300	500	1000
F	54.1	49.9	48.8	53.6	54.6

Table 2. Average F-score on the development set with different number of clusters

3.2. Experimental results

Complete experimental results for the three methods are presented in Table 1, compared with the baseline system in [14].

Full spectrum and complete dictionary. In testing, all the dictionary components obtained in the training are concatenated to form the complete dictionary. The effective size of the dictionary $\mathbf{W}^{(1)}$ in testing is $F \times D_{full}$, where D_{full} is the total number of dictionary components obtained during training. Average event detection accuracy of this method is 55.6%.

Full spectrum and clustered dictionary. To select the best number of clusters for reducing the number of components, we used a development set containing recordings from all contexts. The event detection accuracy was evaluated for a number of 100, 300, 500 and 1000 components, obtained as the centroids of clusters resulting from k-means clustering. Based on the results presented in Table 2, we chose to use 500 clusters for the full set testing. The effective size of the dictionary in testing is $F \times 500$, significantly smaller than for the original method. Average accuracy of this method is 57.8%.

Mel spectrum and complete dictionary. Instead of the full spectrum, we use the 40 band mel magnitude as spectral representation in the NMF algorithm, reducing the dictionary components size from 512 to 40. The effective size of the dictionary $\mathbf{W}^{(1)}$ in testing is $40 \times D_{mel}$, where D_{mel} is the total number of dictionary components obtained during training, which is slightly different than D_{full}. The average event detection accuracy of this method is 56.5%.

Dictionary size for the test phase is significantly reduced for the latter two methods, compared to the original method. The size of the complete dictionary depends on the amount of audio segments available for training. Numerical details of the dictionary size reduction for the used database are presented in Table 3. The mel-magnitude representation offers the same detection accuracy as the original method with a dictionary 16 times smaller, while for clustering, the dictionary that offers a similar performance is 7 times smaller.

context	baseline	full spectrum complete dict	full spectrum clustered dict	mel spectrum complete dict
basketball	55.3	**74.9**	71.7	69.1
beach	33.6	58.9	56.5	**61.5**
bus	**45.5**	30.4	37.0	38.4
car	32.1	33.3	**35.9**	27.5
hallway	46.3	**60.2**	59.3	57.8
office	37.3	47.2	**63.2**	55.2
restaurant	56.2	79.5	77.4	**81.6**
shop	43.8	48.9	**57.9**	51.6
street	48.6	**57.3**	54.6	54.8
tracknfield	49.9	65.3	64.0	**67.7**
overall	44.9	55.6	**57.8**	56.5

Table 1. Event detection accuracy in one second blocks compared to the baseline system [14]

context	clustered dictionary	mel spectrum dictionary
basketball	0.13	0.06
beach	0.06	0.06
bus	0.17	0.08
car	0.28	0.07
hallway	0.17	0.06
office	0.14	0.05
restaurant	0.11	0.05
shop	0.12	0.05
street	0.13	0.05
tracknfield	0.10	0.05
average	0.14	0.06

Table 3. Dictionary size in percentages reported to the size of the complete dictionary of full spectrum

4. DISCUSSION

We compare the performance of our methods with the system presented in [14], using as a baseline the stream selection procedure that resulted in the higher performance. The system in [14] used a supervised classification approach and constructed hidden Markov models for each class, based on components separated from the mixture signal with unsupervised NMF. Our method uses the separated components as such, without modeling classes. This seems to be more robust to the complexity of the audio data and to possible variability in event classes. Table 1 provides direct comparison of the methods performance context by context. The proposed methods outperform the baseline system in all but one context. On average, the methods proposed in this paper have 10% higher performance than the baseline.

The performance of the three proposed methods is very similar – the reduction of the dictionary size does not seem to affect significantly the event detection performance. Reducing the number of components by clustering has the major advantage of giving exact control of the dictionary size. The new dictionary seems to have good generalization properties, but its performance varies both ways – this means that optimizing the number of clusters separately for each context might increase overall performance.

Using the mel spectrum has the major advantage of allowing use of the complete dictionary, as the major size reduction here results from using only 40 mel bands instead of all FFT bins. The mel representation of the spectrum results into good representation in terms of dictionary. Further development as logical step would be to use both mel spectrum and clustering of components, and might offer an efficient way of dealing with very large databases.

5. CONCLUSIONS

This paper presented a method for detection of overlapping events in large database without constructing separate models for each class. The method has the advantage of training directly on complex audio with annotated overlapping sound events, and outputs an annotation matrix containing overlapping sound events. The mid-level representation of the data consists in non-negative dictionary components that are not associated to any sound sources, in this sense the method being unsupervised. Large amounts of training data result in very large dictionary, therefore two ways of reducing the dictionary size were proposed: one reducing the number of components, other reducing the size of the components. Both dictionary reduction methods have similar performance with the original method, while all three have significantly higher performance than the baseline system.

Future work will include methods for controlling the size of the dictionary already in the training phase. This can be achieved for example by using iterative methods for learning the dictionary, to avoid creating redundant components. This would also allow adding training data by simply performing additional training iterations.

6. REFERENCES

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[2] S. Ntalampiras and I. Potamitis, "Detection of human activities in natural environments based on their acoustic emissions," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 1469–1473.

[3] 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.

[4] R. Cai, Lie Lu, A. Hanjalic, H-J. Zhang, and L-H. Cai, "A flexible framework for key audio effects detection and auditory context inference," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 1026 – 1039, may 2006.

[5] M. Xu, C. Xu, L. Duan, Jesse S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 4, no. 2, pp. 11:1–11:23, May 2008.

[6] Huy Dat Tran and Haizhou Li, "Sound event recognition with probabilistic distance svms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1556–1568, Aug 2011.

[7] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M.D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct 2013, pp. 1–4.

[8] A. Temko, C. Nadeu, and J. Biel, "Acoustic event detection: SVM-based system and evaluation setup in clear'07," Berlin, Heidelberg, 2008, pp. 354–363, Springer-Verlag.

[9] A. Temko and C. Nadeu, "Acoustic event detection in a meeting-room environment," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[10] 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.

[11] T. Butko, F. Gonzalez Pla, C. Segura, C. Nadeu, and J. Hernando, "Two-source acoustic event detection and localization: online implementation in a smart-room," in *19th European Signal Processing Conference (EUSIPCO 2011)*, 2011, pp. 1307–1311.

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

[13] 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.

[14] 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.

[15] O. Dikmen and A Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct 2013, pp. 1–4.

[16] J.F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based nmf approach to audio event detection," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct 2013, pp. 1–4.

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

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

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

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

[21] 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.