Detection of Overlapping Acoustic Events using a Temporally-Constrained Probabilistic Model

Emmanouil Benetos, Grégoire Lafay, Mathieu Lagrange, Mark D. Plumbley

DETECTION OF OVERLAPPING ACOUSTIC EVENTS USING A TEMPORALLY-CONSTRAINED PROBABILISTIC MODEL

Emmanouil Benetos[1]*, Grégoire Lafay*[2]*, Mathieu Lagrange*[2]*, and Mark D. Plumbley*[3]

[1] Centre for Digital Music, Queen Mary University of London, UK
[2] IRCCYN, CNRS, École Centrale de Nantes, France
[3] Centre for Vision, Speech and Signal Processing, University of Surrey, UK

ABSTRACT

In this paper, a system for overlapping acoustic event detection is proposed, which models the temporal evolution of sound events. The system is based on probabilistic latent component analysis, supporting the use of a sound event dictionary where each exemplar consists of a succession of spectral templates. The temporal succession of the templates is controlled through event class-wise Hidden Markov Models (HMMs). As input time/frequency representation, the Equivalent Rectangular Bandwidth (ERB) spectrogram is used. Experiments are carried out on polyphonic datasets of office sounds generated using an acoustic scene synthesizer - simulator, as well as real and synthesized monophonic datasets for comparative purposes. Results show that the proposed system outperforms several state-of-the-art methods for overlapping acoustic event detection on the same task, using both frame-based and event-based metrics, and is robust to varying event density and noise levels.

Index Terms— Acoustic event detection, probabilistic latent component analysis, hidden Markov models

1. INTRODUCTION

Acoustic event detection, also called sound event detection, is a central topic in the emerging field of acoustic scene analysis. The main goal of the aforementioned task is to label temporal regions within an audio recording, resulting in a symbolic description with start and end times, as well as labels for each instance of a specific event type [1]. Applications for acoustic event detection are numerous, including but not limited to security and surveillance, urban planning, smart homes, acoustic ecology, and organisation/navigation of sound archives [1, 2, 3, 4].

The majority of research in acoustic event detection is directed towards detecting only one event at a given time segment, which is also referred to as *monophonic* event detection, or detection of non-overlapping acoustic events. Methods that address the problem of detecting overlapping events from audio (also called *polyphonic* event detection) include the work by Heittola et al. on using a context-dependent Hidden Markov Model (HMM) with multiple path decoding [3]. Gemmeke et al. proposed the use of using vectorized time-frequency patches of pre-extracted isolated events within the context of non-negative matrix factorization (NMF) [5]. Dennis et al. [4] detect overlapping sound events using local spectrogram features and a Generalised Hough Transform voting system. A more recent approach for event detection uses multilabel deep neural networks with spectral features as inputs [6]. In addition, as part of the IEEE AASP challenge on Detection and Classification of acoustic scenes and events (DCASE) [7], a baseline system was created using NMF with beta-divergence. Finally, also part of the DCASE challenge, Vuegen et al. [8] proposed a system based on Gaussian mixture models (GMMs), with MFCCs as input features.

In this paper, a method for polyphonic event detection is proposed, based on probabilistic latent component analysis (PLCA - the probabilistic counterpart of NMF). The proposed event detection system is adapted from [9], which was created for automatic music transcription. Here, a dictionary of pre-extracted events is created, which expresses each exemplar as a succession of spectral templates. Temporal constraints modelling the evolution of each produced sound event are incorporated in the proposed model, using event-wise HMMs. Experiments are carried out using the polyphonic event detection dataset from the DCASE challenge, generated using isolated sounds recorded at Queen Mary University of London, as well as a new dataset generated using isolated sounds from IRCCYN, France, in order to test the proposed method's generalization capabilities. Comparative experiments are also made using real and synthesized monophonic datasets. Results show that the system outperforms several state-of-the-art methods for detecting overlapping events, using several types of evaluation metrics.

On relation to prior work: in contrast with the NMF-based systems of [5, 10], which model events using either vectorized or 2-dimensional time-frequency patches, events are here modelled as a temporal succession of spectral templates, leading to a computationally efficient model. Also, in contrast with [11], this model proposes an event class-exemplar-sound state hierarchy, which expresses a test event as a linear combination of exemplars for that specific event class.

The outline of this paper is as follows. The proposed system is presented in Section 2. Evaluation, including a description of the train/test datasets, evaluation metrics, and results, is presented in Section 3. The paper concludes with a discussion in Section 4.

2. PROPOSED SYSTEM

2.1. Motivation

The overall aim of the proposed work is the creation a computationally efficient dictionary-based system for overlapping acoustic event detection that expresses a sound event as a combination of exemplars. Each exemplar consists of a series of spectral templates each corresponding to a *sound state*, the order of which is controlled using temporal constraints. Thus, the model is able to exploit spectro-temporal features without resorting to computationally expensive convolutional formulations [12, 10] or vectorized time-

EB is supported by a Royal Academy of Engineering Research Fellowship (grant no. RF/128).

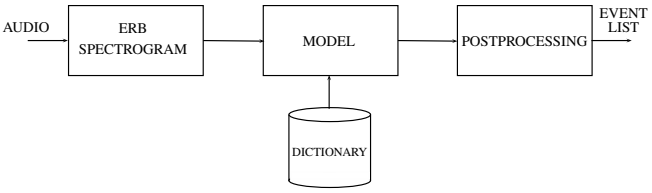

Fig. 1. Proposed system diagram.

frequency patches [13].

The proposed system adapts the model of [9], which was proposed for the task of automatic music transcription, and is based on PLCA [14], a spectrogram factorization method which supports the detection of overlapping sound events. In contrast to [9], the proposed model does not support shift-invariance across a log-frequency representation, as the concept of musical tuning does not apply to everyday sounds. Instead, as input time-frequency representation, we propose the use of the Equivalent Rectangular Bandwidth (ERB) spectrogram [15], which provides a more compact representation compared to the STFT spectrogram. A diagram for the proposed system is shown in Fig. 1.

2.2. Model

The proposed model takes as input a normalised time-frequency representation $V_{f,t}$ (f is the frequency index and t is the time index) and approximates it as a bivariate probability distribution $P(f,t)$. In this work, $V_{f,t}$ is created by processing the input signal with an Equivalent Rectangular Bandwidth (ERB) filterbank [15]. This auditory-motivated filterbank uses 250 filters, linearly spaced between 5Hz and 10.8kHz on the ERB scale, and is computed using the method of [16], where each subband is partitioned into disjoint 23ms time frames, and the rms is computed for each frame. A linear pre-emphasis filter is applied to $V_{f,t}$ in order to boost high frequencies, which in the case of sound event detection carry useful information.

The model decomposes $P(f,t)$ into a series of spectral templates per event class, exemplar index, and sound state, as well as probability distributions for event activations, exemplar contributions per class, and sound state activations per class. The model is formulated as:

$$P(f,t) = P(t) \sum_{q,c,s} P(f|q,c,s)P(s|t)P(c|s,t)P(q|s,t) \quad (1)$$

where s denotes the sound event class, c denotes the exemplar index, and q the sound state index. $P(t)$ is defined as $\sum_f V_{f,t}$, which is a known quantity. $P(f|q,c,s)$ is a 4-dimensional tensor that contains the pre-extracted spectral templates for event s, exemplar c and sound state q. $P(s|t)$ is the time-varying event activation (which is the main output used for evaluation). $P(c|s,t)$ denotes the time-varying exemplar contribution for producing a specific event. Finally, $P(q|s,t)$ is the sound state activation per event class, across time.

$P(s|t)$ and $P(c|s,t)$ can be estimated using iterative update rules through the Expectation-Maximization (EM) algorithm [17]. For the *E-step*, the following posterior is computed:

$$P(q,c,s|f,t) = \frac{P(f|q,c,s)P(s|t)P(c|s,t)P(q|s,t)}{\sum_{q,c,s} P(f|q,c,s)P(s|t)P(c|s,t)P(q|s,t)} \quad (2)$$

For the *M-step*, $P(s|t)$ and $P(c|s,t)$ are updated using the posterior of (2):

$$P(s|t) = \frac{\sum_{q,c,f} P(q,c,s|f,t)V_{f,t}}{\sum_{s,q,c,f} P(q,c,s|f,t)V_{f,t}} \quad (3)$$

$$P(c|s,t) = \frac{\sum_{q,f} P(q,c,s|f,t)V_{f,t}}{\sum_{c,q,f} P(q,c,s|f,t)V_{f,t}} \quad (4)$$

2.3. Temporal constraints

Without any temporal constraints, $P(q|s,t)$ can be estimated using an iterative update rule similar to (3) or (4). In this system however, temporal constraints on the order of the sound states are introduced through the use of hidden Markov models (HMMs). One HMM is created per event class, which has the sound states q as hidden states. Thus, the basic elements of the event-wise HMMs are: the sound state priors $P(q_1^{(s)})$, the sound state transitions $P(q_{t+1}^{(s)}|q_t^{(s)})$ and the observations $P_t(\bar{f}_t|q_t^{(s)})$. Here, $\bar{f}$ corresponds to the sequence of observed spectra from all time frames, $\bar{f}_t$ is the observed spectrum at the t-th time frame, and $q_t^{(s)}$ is the value of the hidden state at the t-th time frame.

On linking the event-wise HMMs with the model of (1), the following assumption is made:

$$P(q|s=i,t) = P_t(q_t^{s=i}|\bar{f}) \quad (5)$$

thus, the sound state activations per event are assumed to be produced by the posteriors of the HMM corresponding to event i. Following [18], the observation probability is calculated as:

$$P(\bar{f}_t|q_t^{(s)}) = \prod_{f_t} P(f_t|q_t^{(s)})^{V_{f,t}} \quad (6)$$

where $P(f_t|q_t^{(s)})$ is the approximated spectrum for a given sound state and event class. This is because in PLCA-based models $V_{f,t}$ represents the number of times f has been drawn at the t-th time frame [18].

For estimating the unknown HMM parameters, the EM algorithm is again used. For the *E-step*, the HMM posterior per event class is computed as:

$$P_t(q_t^{(s)}|\bar{f}) = \frac{P_t(\bar{f}, q_t^{(s)})}{\sum_{q_t^{(s)}} P_t(\bar{f}, q_t^{(s)})} = \frac{\alpha_t(q_t^{(s)})\beta_t(q_t^{(s)})}{\sum_{q_t^{(s)}} \alpha_t(q_t^{(s)})\beta_t(q_t^{(s)})} \quad (7)$$

where $\alpha_t(q_t^{(s)})$ and $\beta_t(q_t^{(s)})$ are the forward and backward variables for the s-th HMM, respectively, and can be estimated using the forward-backward algorithm [19]. The posterior for the sound state transitions $P_t(q_{t+1}^{(s)}, q_t^{(s)}|\bar{f})$ is computed as in [18].

In the *M-step*, the sound state priors and transitions per event class are estimated from the posterior of (7):

$$P(q_1^{(s)}) = P_1(q_1^{(s)}|\bar{f}) \quad (8)$$

$$P(q_{t+1}^{(s)}|q_t^{(s)}) = \frac{\sum_t P_t(q_t^{(s)}, q_{t+1}^{(s)}|\bar{f})}{\sum_{q_{t+1}^{(s)}} \sum_t P_t(q_t^{(s)}, q_{t+1}^{(s)}|\bar{f})} \quad (9)$$

Overall, for estimating all unknown parameters ($P(s|t)$, $P(c|s,t)$, $P(q|s,t)$), both the PLCA-based and the HMM-based update rules are used in an iterative fashion. For the *E-step*, the model posterior is computed using (2), followed by the HMM posteriors of (7). For the *M-step*, $P(s|t)$ and $P(c|s,t)$ are estimated using (3) and (4), respectively. $P(q_1^{(s)})$ and $P(q_{t+1}^{(s)}|q_t^{(s)})$ are estimated using (8) and (9), respectively. Finally, $P(q|s,t)$ is estimated using (5). For this implementation, the algorithm was set to 30 iterations.

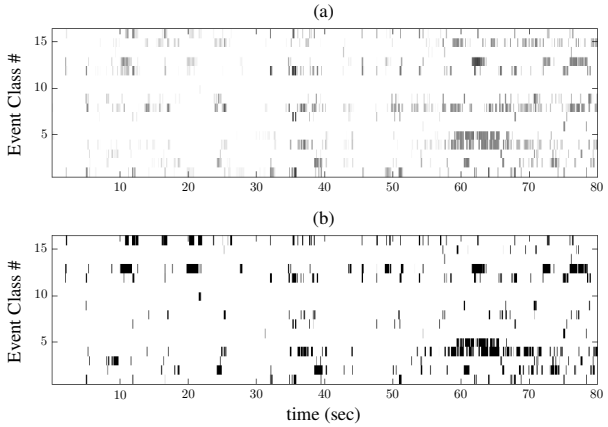

Fig. 2. (a) The event activation $P(s,t)$ for a recording of the DCASE OS test dataset. (b) The post-processed event-roll. Class IDs 1-16 are described in Sec. 3.1.

2.4. Post-processing

The output of the proposed model is the event activation, weighted by $P(t)$: $P(s,t) = P(t)P(s|t)$. This is a non-binary representation, which needs to be converted into a list of detected events per time frame. Here, $P(s,t)$ is post-processed by performing median filtering across time, (with an 180ms span), followed by thresholding (values are estimated using a development set, cf. subsection 3.1). Finally, events with a small duration (shorter than 60ms) are removed. Fig. 2 shows an example event activation, along with the post-processed binary event-roll which is used for evaluation.

3. EVALUATION

3.1. Training Data

For constructing the pre-extracted dictionary $P(f|q,c,s)$, the IEEE DCASE Event Detection training dataset was used [7, 1]. The dataset contains isolated sounds recorded in an office environment at Queen Mary University of London, and covers 16 event classes ($s \in \{1,...,16\}$): alert, clearing throat, cough, door slam, drawer, keyboard click, keys, door knock, laughter, mouse click, page turn, pen drop, phone, printer, speech, and switch. Each class contains 20 exemplars ($c \in \{1,...,20\}$). In this work, the number of sound states was set to 3 ($q \in \{1,2,3\}$) following experimentation. In order to extract the sound state templates, each isolated sound ERB spectrogram was split into 3 segments with equal duration. PLCA with a single component was applied to each segment in order to extract a single sound state spectral template. For tuning system parameters for the polyphonic and monophonic datasets, the development datasets for the IEEE DCASE Office Synthetic and Office Live challenge [7] were used, respectively.

3.2. Test Data

For testing, 2 polyphonic datasets of artificially concatenated office sounds were used, with varying levels of event density (i.e. polyphony) and SNR. In addition, 3 monophonic datasets (1 real and 2 synthesized) of office sounds were also used, for comparative purposes.

On the polyphonic datasets, firstly the test dataset for the IEEE DCASE Office Synthetic (OS) challenge was used [1]. The dataset contains 12 recordings of 2min duration each, with 3 different event density levels (low, mid, high) and 3 different SNR levels (-6dB, 0dB, and 6dB). The recordings were generated by concatenating isolated office sounds recorded at Queen Mary University of London (using different sources than the ones used for the training dataset of subsection 3.1), using the acoustic scene synthesizer of [20]. This dataset allows for a detailed evaluation wrt the proposed method's capabilities on different polyphony and background noise levels. The second polyphonic dataset uses the same event ground-truth with the OS dataset, as well as the same noise level and event density settings, but was instead generated using samples recorded at the École Centrale de Nantes, France. This second dataset, which will be called OS-IRCCYN dataset from now on, thus allows for evaluating the proposed method's generalization capabilities.

For comparative purposes, 3 monophonic datasets of office sounds were used. Firstly, the Office Live (OL) dataset from the DCASE challenge was used [1], which contains 11 scripted recordings of event sequences recorded at Queen Mary University of London. The second and third monophonic datasets were generated using the acoustic scene synthesizer of [20], and each include 22 recordings of variable duration (1-3min), using as basis the annotations for the OL dataset. Both synthesized datasets were generated using isolated sounds recorded at the École Centrale de Nantes, France, thus are useful for testing the proposed method's generalization capabilities. The second monophonic dataset was generated using the *instance* simulation process, while the third dataset was generated using the *abstract* simulation process (see [20] for more details).

	$\mathcal{F}_f$	$\mathcal{F}_{eb}$	$\mathcal{F}_{cweb}$
Stowell et al. [1]	12.8%	7.8%	9.5%
Vuegen et al. [8]	13.5%	13.8%	10.5%
Heittola et al. [3]	18.7%	16.1%	18.7%
Gemmeke et al. [5]	21.3%	17.0%	14.2%
Proposed System	25.6%	21.8%	20.6%

Table 1. Event detection results for the polyphonic DCASE OS test dataset.

3.3. Metrics

For evaluation, the same set of metrics used for the IEEE DCASE event detection tasks was used [1]. Specifically, 3 different metrics are used: frame-based, event-based, and class-wise event-based. Frame-based evaluation is performed on a 10ms step using the post-processed event activation, while event-based and class-wise event-based evaluation consider each event to be correctly detected if its onset is within a $\pm$100ms onset tolerance. In all cases, the F-measure is reported ($\mathcal{F}_f$, $\mathcal{F}_{eb}$, and $\mathcal{F}_{cweb}$, respectively).

3.4. Results

This subsection presents evaluation results using the proposed method, as well as comparisons using the publicly available systems of [5] (using a frame stacking-based NMF approach), [8] (using an MFCC-GMM approach), as well as the NMF-based baseline system for the IEEE DCASE challenge [1]. For the 1st test dataset, results are also shown for the HMM-based multiple path decoding system of [3], as reported in the DCASE challenge.

Event detection results using the OS test dataset for the proposed method and the aforementioned comparative systems are presented in Table 1, averaged across all recordings. It can be seen that the

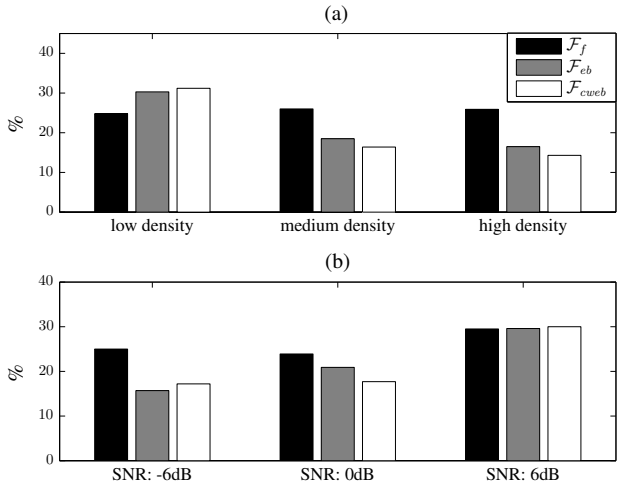

Fig. 3. Event detection results for the proposed system on the polyphonic DCASE OS dataset for (a) varying event density (b) varying SNR levels.

	$\mathcal{F}_f$	$\mathcal{F}_{eb}$	$\mathcal{F}_{cweb}$
Stowell et al. [1]	13.8%	2.9%	2.4%
Vuegen et al. [8]	3.5%	1.1%	4.7%
Gemmeke et al. [5]	10.8%	5.9%	3.3%
Proposed System	14.7%	6.4%	5.6%

Table 2. Event detection results for the polyphonic OS-IRCCYN test dataset.

proposed method outperforms all comparative approaches using the 3 different metrics. It is worth pointing out that frame-based metrics are generally larger than event-based metrics, which shows that more work can be done in grouping frame activations into contiguous events with a start and end time. Additional information on the proposed method's performance is presented in Fig. 3, showing results for groups of recordings with specific polyphony and SNR levels. It is worth noting that frame-based performance is stable across different polyphony levels, whereas event-based metrics drop with increased event density. On SNR levels, again frame-based metrics are more stable compared to event-based metrics with increased noise level. This shows that while the system can detect events irrespective of density and noise levels, tracking and grouping events in noisy multisource environments would require an alternate approach to the one presented in Section 2.4.

Results using the OS-IRCCYN dataset are shown in Table 2, for the proposed and comparative methods, averaged across all recordings. A significant drop compared to the OS dataset results can be seen across all methods, which can be attributed to the different recording equipment and conditions used to generate the isolated samples compared to the OS dataset. In particular, a significant drop is reported for [8], while the baseline system of [1] is relatively robust. The proposed method ranks best across all metrics, although the results clearly indicate that source- and recording condition-independent polyphonic event detection is a problem that would need to be addressed in the future.

Comparative results on the proposed method's performance for monophonic event detection are shown in Table 3, using the Office Live dataset, as well as the synthesized *instance* and *abstract* datasets using samples from IRCCYN. For the OL dataset, the method of [8] ranks best across all metrics, followed by the proposed method (it should be noted that [8] was also trained on the OL development dataset). This changes when using the IRCCYN-generated monophonic sequences, where the proposed method shows better generalization capabilities across all metrics. Still, there is a significant gap between the performance of frame-based metrics and event-based metrics, which again indicates that aggregating frame-based detections for forming coherent events requires a methodology that goes beyond thresholding and minimum duration pruning.

	$\mathcal{F}_f$	$\mathcal{F}_{eb}$	$\mathcal{F}_{cweb}$
Vuegen et al. [8]	43.4%	30.8%	24.6%
Stowell et al. [1]	10.7%	7.4%	9.0%
Gemmeke et al. [5]	31.9%	15.5%	13.2%
Proposed System	34.4%	23.8%	21.0%
Vuegen et al. [8]	9.5%	9.3%	7.3%
Stowell et al. [1]	14.0%	6.4%	5.7%
Gemmeke et al. [5]	18.5%	11.2%	6.0%
Proposed System	23.6%	16.3%	11.4%
Vuegen et al. [8]	9.7%	10.2%	7.3%
Stowell et al. [1]	14.4%	5.9%	5.6%
Gemmeke et al. [5]	18.8%	9.2%	5.4%
Proposed System	26.8%	14.4%	11.4%

Table 3. Monophonic event detection results for the OL (top group), *instance* (middle group) and *abstract* (bottom group) datasets [20].

Finally, regarding runtimes, the proposed method performs at about 0.7 $\times$ real-time using a (3-year old) Sony VAIO S15 laptop using a Matlab implementation. This shows that the proposed approach can be used in applications requiring computational efficiency, such as for real-time event detection.

4. DISCUSSION

This paper proposed a computationally efficient method for polyphonic acoustic event detection based on an HMM-constrained PLCA-based model with an event class-exemplar-sound state hierarchy. As input time-frequency representation, the ERB spectrogram was used. Experiments on both polyphonic and monophonic datasets of office sounds showed that the proposed method outperforms other approaches in the literature.

However, results also show that the problem of overlapping event detection is still far from being solved. Of particular importance for future research would be the adaptation of event detection systems to various recording environments and conditions, different sound sources, and variable noise levels. In addition, the template extraction process of section 3.1 will be revised, as to use a different number of sound states per event class. Another issue in need to be resolved would be the discrepancy between frame-based metrics and event-based metrics, which shows that an additional post-processing step is needed in order to track and form coherent events across time. To that end, future work will focus on a *transduction* post-processing step for converting the non-binary event activation into a list of events with start and end times, using as basis probabilistic machine learning methods for sequential data.

5. REFERENCES

[1] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[2] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 321–329, 2006.

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

[4] J. Dennis, H.D. Tran, and E.S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1085–1093, 2013.

[5] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audio event detection," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013.

[6] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *International Joint Conference on Neural Networks*, July 2015.

[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 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, Oct. 2013.

[8] L. Vuegen, B. Van Den Broeck, P. Karsmakers, J. F. Gemmeke, B. Vanrumste, and H. Van hamme, "An MFCC-GMM approach for event detection and classification," *IEEE AASP DCASE Challenge*, 2013, `http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/abstracts/OS/VVK.pdf`.

[9] E. Benetos and T. Weyde, "An efficient temporally-constrained probabilistic model for multiple-instrument music transcription," in *16th International Society for Music Information Retrieval Conference (ISMIR)*, Malaga, Spain, October 2015.

[10] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2011, pp. 69–72.

[11] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," in *European Signal Processing Conference*, Barcelona, Spain, 2011, pp. 1307–1311.

[12] E. Benetos, M. Lagrange, and S. Dixon, "Characterisation of acoustic scenes using a temporally-constrained shift-invariant model," in *15th International Conference on Digital Audio Effects (DAFx)*, York, UK, Sept. 2012, pp. 317–323.

[13] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[14] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizations," *Computational Intelligence and Neuroscience*, 2008, Article ID 947438.

[15] B. C. J. Moore, "Frequency analysis and masking," in *Hearing – Handbook of Perception and Cognition*, B. C. J. Moore, Ed., pp. 161–205. Academic Press, San Diego, CA, 2nd edition, 1995.

[16] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[18] G. Mysore, *A Non-negative Framework for Joint Modeling of Spectral Structure and Temporal Dynamics in Sound Mixtures*, Ph.D. thesis, Stanford University, USA, June 2010.

[19] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[20] M. Lagrange, G. Lafay, M. Rossignol, E. Benetos, and A. Roebel, "An evaluation framework for event detection using a morphological model of acoustic scenes," *ArXiv e-prints*, Jan. 2015, arXiv:1502.00141 [stat.ML].