AN EXEMPLAR-BASED NMF APPROACH FOR AUDIO EVENT DETECTION

Jort F. Gemmeke[1], *Lode Vuegen*[1,2,3], *B. Vanrumste*[2,3,4], *H. Van hamme*[1]

[1]ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium
[2]iMinds, Future Health Department, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium
[3]MOBILAB, TM Kempen, Kleinhoefstraat 4, 2440, Geel, Belgium
[4]ESAT-SISTA, KU Leuven, Kasteelpark Arenberg 10, 3001, Heverlee, Belgium

email: jgemmeke@amadana.nl

ABSTRACT

We present a novel, exemplar-based method for audio event detection based on non-negative matrix factorisation (NMF). Building on recent work in noise robust automatic speech recognition, we model events as a linear combination of dictionary atoms, and mixtures as a linear combination of overlapping events. The exemplar-based dictionary is created by extracting all available training data, artificially augmented by linear time warping at multiple rates. The method is evaluated on the Office Live and Office Synthetic development datasets released by the AASP Challenge on Detection and Classification of Acoustic Scenes and Events.

Index Terms— Audio event detection, exemplars, NMF

1. INTRODUCTION

Automatic *audio event detection* is an application of pattern recognition and machine learning in which an audio signal is mapped to a symbolic description of the corresponding sound event(s) present in the auditory scene. Automatic audio event detection is utilized in a host of applications, including context-based indexing and retrieval in multimedia such as movies and sports videos, unobtrusive monitoring in health care, surveillance, lifeblogging, audio segmentation and military applications.

Most conventional audio event detection techniques employ Gaussian Mixture Models (GMMs), operating on Mel-Cepstral Coefficients (MFCCs) [1]. In this work, we built on state-of-the-art noise robust Automatic Speech Recognition (ASR) techniques [2, 3] and pursue an exemplar-based Non-negative Matrix Factorisation (NMF) approach to audio event detection. NMF has only recently been considered in the context of audio event detection [4, 5, 6] and our submission offers three contributions: 1) modelling all training data through *exemplars* which facilitates the use of long temporal contexts and large, possibly overcomplete dictionaries, 2) artificially increasing the (limited) amount of training data through resampling at various rates, and 3) explicit modelling of background events such as noise. The method is evaluated on the Office Live and Office Synthetic development datasets released by the AASP Challenge on Detection and Classification of Acoustic Scenes and Events [7].

This research was funded by the IWT-SBO project ALADIN (contract 100049) and an IWT doctoral scholarship (contract 121565).

2. METHOD

2.1. Compositional model

The compositional model for audio event detection is based on representing both individual audio events, as well as mixtures of audio events, as a linear combination of atoms. The collection of audio event atoms form a *dictionary*, and in this work atoms are formed by *exemplars*, spectrogram segments extracted from a set of training samples.

The atoms are $B \times T$ magnitude spectrogram segments, reshaped to a $E = B \cdot T$ dimensional vector, where B is the number of frequency bands and T is the number of consecutive time frames in an atom. With an observed signal $\mathbf{\Psi}$ represented in the form of overlapping windows of length T, a spectrogram window $\mathbf{Y}$ is reshaped to an E-dimensional vector $\mathbf{y}$ and approximated as:

$$\mathbf{y} \approx \left[\mathbf{A}^1 \mathbf{A}^2 \cdots \mathbf{A}^D\right] \begin{bmatrix} \mathbf{x}^1 \\ \mathbf{x}^2 \\ \vdots \\ \mathbf{x}^D \end{bmatrix} = \mathbf{A}\mathbf{x} \quad \text{s.t.} \quad \mathbf{x} \geq 0 \qquad (1)$$

with D the total number of audio event dictionaries, and $\mathbf{x}^d$ the weight of the linear combination of atoms in dictionary $\mathbf{A}^d$, $1 \leq d \leq D$. Each dictionary $\mathbf{A}^d$ is a matrix of size $E \times N^d$, with N the number of atoms. The total number of atoms in $\mathbf{A}$ is M.

The model (1) can be interpreted as an instance of non-negative matrix factorisation (NMF), with a fixed dictionary $\mathbf{A}$ that is determined in advance. For all window spectrograms in a observed signal, the representation $\mathbf{x}$ is obtained by solving a convex optimisation problem. The cost function to be minimized is composed of the Kullback-Leibler divergence between $\mathbf{y}$ and $\mathbf{A}\mathbf{x}$ augmented with a sparsity inducing term composed of the sum of $\mathbf{x}$ entries weighted a penalty λ. For details on this cost function and its optimisation we refer the reader to [3] and the references therein.

2.2. Dictionary creation

In the exemplar-based approach presented in [3], spectrograms underlying dictionary atoms are directly extracted from training samples pertaining a particular source. Since in that work, the amount of training data is rather large, random sampling was used. This abstract, however, details work on the AASP audio event challenge which has only a limited amount of training data. Therefore, for each audio event a dictionary is composed by exhaustively sampling all possible, full-length non-overlapping spectrogram windows from each of the samples available for that event. If a sample

does not even span a single window (its duration is shorter than T frames), which happens for some short events such as "switch", the spectrogram is zero-padded up to T frames.

While the use of atoms that span multiple time frames has proven to be advantageous in noise robust ASR [3], a disadvantage is that the explicit modelling of time context makes it difficult to match observations with (local) durations that do not quite match those in the limited amount of training data. In order to circumvent this data scarcity issue, we artificially increase the amount of training data for short (less than 5 seconds) samples by linear temporal warping of the spectrograms at the rates: 0.75, 1.25 and 1.5.

In this work, we treat background (both non-target events and ambient noise) as its own audio event: The annotation of the training samples is used to determine the start and end point of the event by taking the maximum possible duration (over both annotators). The event data is used for dictionary extraction as described above - the non-event data before and after the event is concatenated and similarly used to extract background dictionary atoms. Additionally, during decoding a small number of atoms are extracted from the begin and end of the observed signal and added to the background dictionary on-the-fly [2].

2.3. Audio event detection

In correspondence with earlier work on ASR [8], we introduce a label-atom mapping $\mathbf{L}$ to associate atoms to events. $\mathbf{L}$ is a $D \times M$ dimensional binary matrix, with a non-zero entry in the d-th row indicating that a certain atom is associated with audio event d. For each sliding window position in the observed signal we estimate the (unscaled) presence of events as:

$$\mathbf{o} = \mathbf{L}\mathbf{x} \tag{2}$$

with $\mathbf{o}$ a D-dimensional vector indicating the activation ('likelihood') of events.

The event-likelihood matrix $\mathbf{O}$ describing the entire observed signal $\mathbf{\Psi}$ is formed by overlap-adding the sliding window estimates $\mathbf{o}$, under the assumption that an activated audio event spans the entire duration T. $\mathbf{O}$ was converted to posterior probability estimates in three steps: First, $\mathbf{O}$ was scaled through division by its largest entry [4, 8]. Second, a small 'background offset' s was added to the likelihoods pertaining the background event. This can be beneficial since in the absence of any sound, the representation $\mathbf{x}$ approaches zero for all atoms and hence $\mathbf{o}$ becomes all-zero [8]. In the final, third step the entries of the posterior probability estimates were column-wise normalised to sum to one.

For the Office Live dataset, we additionally smoothed the posterior probability estimates using a moving average sliding window filter, with a event-dependent duration. For foreground events, the duration was set to a third of the minimum duration of the events in the training data, averaged over both annotators and all samples, with a maximum of 100 frames. The background event was not filtered. For the Office Synthetic dataset, filtering was not used due to the widely varying acoustic densities of the acoustic scenes.

These event probability estimates were then processed using a Hidden Markov Model (HMM) consisting of a single state per event. An event could transition to any other event, governed by the self-transition probability p_{st}, and the foreground-event-to-background-event transition probability $p_{\mathrm{fb}} = f_{\mathrm{fb}} * (1 - p_{\mathrm{st}})$, with f_{fb} the fraction of non-self-transition probability allocated to transitioning to the background state. The remaining transitions were all equal with the total transition probability summing to one. The most likely sequence of events was determined using the Viterbi algorithm. The Viterbi path was constrained to start and end in a background event.

Table 1: Results on the Office Live dataset for various evaluation methods and metrics. These are averages over both annotators and over all three development files.

Metrics	Evaluation Method		
	Event Based	Class-wise Event Based	Frame Based
R	43.1	38.3	56.4
P	51.7	38.3	77.6
F-score	**46.8**	**36.7**	**65.2**
AEER	1.37	1.18	0.75
Offset R	37.5	31.7	-
Offset P	45.6	33.2	-
Offset F-score	41.0	31.2	-
Offset AEER	1.55	1.38	-

3. EXPERIMENTAL SETTINGS

Acoustic feature vectors consisted of Mel-magnitude spectrograms, spanning $B = 56$ bands. We used the original sampling frequency of 44100 Hz, a pre-emphasis of 0.97, and windowed using a hamming window with a frame length of 25 ms and a frame shift of 10 ms. Stereo data was converted to mono by averaging in the feature domain.

We used exemplars spanning 200 ms, $T = 20$ frames. A sliding window approach was used with windows shifted by a single frame (i.e. 10 ms). The total dictionary size is $M = 10621$, which includes 100 'background' atoms extracted from the first and last 50 frames of the observed signal. Dictionary rows were normalised to equal L-2 norm and dictionary columns were normalised to unit L-2 norm. Optimisation was carried out using 200 iterations of multiplicative updates [3], in single precision. We refer the reader to [3, 2, 8] for algorithmic implementation details.

The parameters that were tuned were λ, s, p_{st} and f_{fb}. The parameters used in this work were tuned on the Office Live dataset by maximizing the average F-score over all metrics, all development files and both annotators. Although we also carried out tuning on the Office Synthetic dataset and for individual metrics and annotators, we found that the (absolute) differences were rather small, in the order of 2-3%. For convenience, and to prevent overfitting, we opted for a single choice of parameter settings for both datasets and for all metrics. We also explored the use of convolutive decoding [4] rather than sliding window based coding, which seemed to improve the results on the Office Synthetic dataset by 2-4%. Again, however, we opted to use a single, sliding window based approach for both datasets.

The sparsity weight was set to $\lambda = 0.3$ for all events. This weight was obtained by a grid search over the range $\lambda \in \{0, 0.1, 0.25, 0.3, 0.4, 0.45, 0.5, 0.75, 1\}$, allowing a different sparsity weight for the foreground and background events during tuning. The background offset was set to $s = 0.008$, after a search over $s \in \{0, 0.0001, 0.0005, 0.001, 0.005, 0.008, 0.01, 0.02\}$. The HMM self-transition probability was set to $p_{\mathrm{st}} = 0.99999$ after a search over $p_{\mathrm{st}} \in \{0.8, 0.9, 0.99, 0.999, 0.9999, 0.99999\}$. Finally, the silence fraction was determined to be $f_{\mathrm{fb}} = 0.001$ after a search over $f_{\mathrm{fb}} \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.4\}$. All parameters were jointly tuned in a five dimensional grid search.

4. EXPERIMENTS ON DEVELOPMENT DATA

4.1. Results

The results on the development data of the Office Live dataset, averaged over the three development files and over both annotators,

Table 2: F-score results on the Office Synthetic Dataset for various evaluation methods. The dataset consists of nine files: three SNRs (-6,0 and 6 dB) at three acoustic 'densities' (low, medium and high).

		Density	SNR [dB]		
			-6	0	6
Evaluation Method	Event Based	low	16.7	10.0	8.7
		medium	5.3	17.0	28.0
		high	4.0	15.5	22.0
	Class-wise Event Based	low	16.7	10.0	6.1
		medium	4.4	14.2	22.2
		high	3.8	12.1	18.7
	Frame Based	low	22.9	24.0	15.9
		medium	37.4	32.7	42.7
		high	21.0	33.4	32.7

are shown in Table 1. The results on the development data of the Office Synthetic dataset are shown in Table 2. For brevity, only the F-scores are shown. The average F-scores over all conditions are 14.1%, 12.0% and 29.2% for the event based, class-wise event based and frame based metrics, respectively.

5. DISCUSSION

5.1. Office live dataset

When comparing the results of the proposed framework on the Office Live dataset (c.f. Table 1) with the baseline results reported in [9], we obtain substantially higher F-scores for all metrics. For example, the baseline system achieves a F-score of 20.6% on the frame-based metric, whereas our exemplar-based framework achieves 65.2%. Since the baseline system also employs an NMF-based framework, it will be interesting to study the commonalities and differences in future work.

A brief study of the underlying confusions showed that most errors were due to brief events such as 'switch'. Also, in the compositional framework some of the events were erroneously modelled by other events with a more parts-based nature: an event with a more or less uniform energy distribution in time energy, would be modelled by (parts of) an event which span only part of the time-frequency spectrum. A possible method to alleviate this would be to use group sparsity [10], which penalizes cross-event atom activations.

The fact that the use of an additional moving average filtering step yields a large improvement in results (in the order of 10% absolute F-score for the frame based metric) serves as indirect evidence that taking typical (minimum) durations of events into account is important. This, and the fact that the self-transition probabilities are tuned to a very high value, indicates that there is substantial room for improvement in modelling the long-term temporal structure in the backend.

5.2. Office synthetic dataset

Overall, we can observe that the performance on the Office Synthetic dataset are much lower than for the Office Live dataset, presumably due to the effect of added noise and overlapping events. Table 2 allows us to study the performance of the proposed framework on the Office Synthetic dataset as a function of SNR and acoustic density. While it should be noted that each reported F-score is solely based on a single recording, we can observe that the highest F-scores for the medium and high density condition are obtained at an SNR of 6 dB. Although the compositional framework can inherently handle the overlap between noise and other events, only a very small set (100) of noise atoms are available, while noise is typically less structured and thus harder to model.

Perhaps surprisingly, the most difficult condition is the 'low' density setting, for which the best results are obtained at the lowest SNR. However, in this condition the acoustic events are so rare that it is likely that these results are not fully representative. Closer inspection revealed that the temporal location of acoustic events was more or less correctly determined, even in high noise conditions, but that the events themselves were incorrectly recognized. A more detailed analysis in future work will have to reveal whether this is an effect of the corrupting noise or due to some other test-train mismatch.

In the 'high' density setting, the results also drop w.r.t the 'medium' density settings. Although the compositional framework can handle overlapping events, the use of a HMM-based decoder as a smoothing step precludes overlapping events. We briefly experimented with multiple Viterbi passes (at each pass zeroing out all event activations of the previous passes), as used in [1], but this did not yield satisfactory results since it led to a large number of insertion errors.

6. CONCLUSIONS AND FUTURE WORK

We presented an exemplar-based NMF framework, which yielded promising results on the Office Synthetic dataset and substantially outperformed the baseline system on the Office Live dataset. Future work will focus on the use of group sparsity to improve acoustic modelling, and the use of more robust back-end models to replace or augment the HMM-based smoothing. Some possibilities are the use of explicit-duration HMMs to model the typical lengths of acoustic events, and the use of more fine-grained state-based models such as those explored in [3].

7. REFERENCES

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

[2] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, "Toward a practical implementation of exemplar-based noise robust ASR," in *Proc. EUSIPCO*, 2011, pp. 1490–1494.

[3] 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.

[4] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *Proc. WASPAA*, 2011, pp. 69–72.

[5] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *CHiME workshop*, 2011.

[6] A. Mesaros, H. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection," in *Proc. EUSIPCO*, 2011.

[7] http://www.elec.qmul.ac.uk/digitalmusic/sceneseventschallenge/.

[8] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. ICASSP*, 2010, pp. 4546–4549.

[9] D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, and M. D. Plumbley, "A database and challenge for acoustic scene classification and event detection," in *submitted to Proc. EUSIPCO*, 2013.

[10] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition," in *Proc. Interspeech*, 2012.