ACOUSTIC EVENT DETECTION FROM WEAKLY LABELED DATA USING AUDITORY SALIENCE

Zuzanna Podwinska[⋆] *Iwona Sobieraj*[†] *Bruno M Fazenda*[⋆] *William J Davies*[⋆] *Mark D. Plumbley*[†]

[⋆] Acoustics Research Centre,
University of Salford, Salford, United Kingdom

[†] Centre for Vision Speech and Signal Processing,
University of Surrey, Guildford, United Kingdom

ABSTRACT

Acoustic Event Detection (AED) is an important task of machine listening which, in recent years, has been addressed using common machine learning methods like Non-negative Matrix Factorization (NMF) or deep learning. However, most of these approaches do not take into consideration the way that human auditory system detects salient sounds. In this work, we propose a method for AED using weakly labeled data that combines a Non-negative Matrix Factorization model with a salience model based on predictive coding in the form of Kalman filters. We show that models of auditory perception, particularly auditory salience, can be successfully incorporated into existing AED methods and improve their performance on rare event detection. We evaluate the method on the Task2 of DCASE2017 Challenge.

Index Terms— AED, NMF, auditory salience, Kalman filter

1. INTRODUCTION

Acoustic Event Detection (AED) is an important task of machine listening, which aims to automatically recognise, label, and estimate the position in time of sound events in a continuous audio signal. In recent years it has seen a rise in interest in the research community, due to the number of real-world applications for AED such as home-care [1], surveillance [2], multimedia retrieval [3] or urban traffic control [4], to name just a few. A successful series of Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [5, 6] have provided the community with datasets and baselines for a number of tasks related to AED, accelerating research in this field. Early approaches for AED were strongly inspired by speech recognition systems, using mel frequency cepstral coefficients (MFCCs) with Gaussian Mixture Models (GMMs) combined with Hidden Markov Models (HMM) [7, 8]. Later, methods based on dictionary learning, mainly Non-negative Matrix Factorization were most prominent solutions for the AED task [9, 10, 11]. In recent years a number of deep learning methods have been proposed, that achieve state of the art results [12, 13]. However, most of the models used for AED nowadays have been primarily developed for image or text processing, while, to our knowledge, the importance of human sound perception in developing methods for AED has not received enough attention. In fact, in many applications of AED it may be desirable to only detect events which human listeners would detect. Therefore, in this paper, we want to pave the way towards perceptually motivated AED, showing how AED models can benefit from the combination with auditory attention and salience models.

Salience is a property of sound which makes it stand out in an auditory scene, and grab listener's attention in a bottom-up manner. There have been a few different approaches to modelling auditory salience, such as calculating salience features from spectrogram images [14], Bayesian surprise [15] and predictive coding [16] (for a review of auditory attention models, see [17]). All of these methods can be interpreted as novelty detection, which makes them natural candidates for an AED task. In fact, [18] has shown that an auditory salience model can be used for AED with promising results. However, these methods have concentrated on detection of every prominent sound, whereas a more realistic scenario might require the detection of a a-priori specified event only, which is the case investigated in this paper.

In this work we propose to combine the best of two worlds and show how they can compliment each other: a Non-negative Matrix Factorization model with an auditory salience model based on Kalman filters. What is more, we show that the resulting method is applicable for AED on *weakly labeled* data, that is, data in which we do not have exact information of when the interesting sound occurs, but just a tag of which sounds are present in a given audio excerpt.

2. AUDITORY SALIENCE MODEL

The auditory salience model used here was developed by [16], and is based on predictive coding, which has been proposed as the working principle of the auditory cortex [19]. The model uses a Kalman filter which: 1) predicts present input based on past input and an internal model, 2) updates the model according to prediction error and 3) weights the model versus data depending on variance of the input. Because salience is related to surprise, understood here as violation of expectation [20], a salient event is detected when the incoming input varies significantly from the prediction.

Each feature extracted from the signal is tracked by one or multiple Kalman filters simultaneously. The state $\mathbf{X}$ of a filter is coded as a vector containing a feature value z_n and the difference between the last two consecutive feature values:

$$\mathbf{X}_n = \begin{bmatrix} z_n \\ z_n - z_{n-1} \end{bmatrix}. \quad (1)$$

At each feature frame n, Kalman filter steps are:

The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement n 642685 MacSeNet. This work is also partly supported by EPSRC grant EP/N014111/1.

$$
\begin{aligned}
\hat{\mathbf{X}}_n^- &= \mathbf{A}\hat{\mathbf{X}}_{n-1} \\
\mathbf{P}_n^- &= \mathbf{A}\mathbf{P}_{n-1}\mathbf{A}^T + \mathbf{Q} \\
\mathbf{K}_n &= \mathbf{P}_n^-\mathbf{B}^T(\mathbf{B}\mathbf{P}_n^-\mathbf{B}^T + \mathbf{R})^{-1} \\
\hat{\mathbf{X}}_n &= \hat{\mathbf{X}}_n^- + \mathbf{K}_n(z_n - \mathbf{B}\hat{\mathbf{X}}_n^-) \\
\mathbf{P}_n &= (\mathbf{I} - \mathbf{K}_n\mathbf{B})\mathbf{P}_n^-
\end{aligned} \tag{2}
$$

where $\hat{\mathbf{X}}^-$ and $\hat{\mathbf{X}}$ are the a priori and a posteriori predictions of $\mathbf{X}$, respectively.

Taken together with the system matrix $\mathbf{A}$ and the measurement matrix $\mathbf{B}$ shown below, this means that at any point in time, the feature vector is expected to continue changing in the same manner it has most recently changed:

$$
\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{B}_n = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{3}
$$

The measurement and system noise covariance matrices are as follows:

$$
\mathbf{Q} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}, \mathbf{R} = \sigma_v^2 \tag{4}
$$

where σ_w, σ_b and σ_v are empirically chosen for each feature.

To determine the number of filters to be initialised at the beginning of each file, Gaussian Mixture Model clustering is performed on the first 500 ms of each feature stream and a Kalman filter is initialised for each of the clusters. The filters are initialised with the same $\mathbf{X}_{n-1}$ and $\mathbf{P}_{n-1}$ values as in [16].

As long as a filter predicts feature values reasonably well, it updates its future predictions based on the input. Feature values are assumed to be predicted well by a filter as long as the following condition is satisfied:

$$
|z_n - \mathbf{B}\hat{\mathbf{X}}_n^-| \leq 2\sqrt{\mathbf{P}_{[1]} + \sigma_v^2} \tag{5}
$$

where $\mathbf{P}_{[1]}$ is the first element of matrix $\mathbf{P}$. When a new feature value is not predicted by any of the currently running filters, a salience spike is produced, with amplitude equal to the difference between prediction and measured value (the system innovation). Additionally, a new filter is initialised based on this new value. If a filter has not correctly predicted any feature values for 1 s, it is closed.

This process produces vectors of salience spikes s_i (one from each feature). The resulting salience score for frame n is obtained by applying feature-specific and between-feature weights and summing the resulting vectors, as follows:

$$
s(n) = \sum_{i\in[1,N]} s_i(n)\left(w_i + \sum_{j\in[1,N],j\neq i} w_{ij} \max_{k\in[-1,1]} s_j(n+k)\right). \tag{6}
$$

3. ORTHOGONALITY-REGULARIZED NMF

For the AED model we use a model from our previous work, where we adapted a standard NMF approach to learning on weakly labeled data [21].

3.1. Non-negative Matrix Factorization

The goal of NMF is to approximate a non-negative data matrix, typically a time-frequency representation of a given sound, $\mathbf{V} \in {R_+}^{F\times T}$ as a product of a dictionary $\mathbf{W} \in {R_+}^{F\times K}$ and its activation matrix $\mathbf{H} \in {R_+}^{K\times T}$, such that:

$$
\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{W}\mathbf{H}. \tag{7}
$$

$\mathbf{W}$ and $\mathbf{H}$ are estimated to minimize some divergence metric $D(\mathbf{V}|\mathbf{W}\mathbf{H})$. For any two matrices X and Y, we define $D(\mathbf{X}|\mathbf{Y}) = \sum_{m,n} D(x_{mn}, y_{mn})$. In this work, we choose the generalized Kullback-Leibler (KL) as the divergence metric, defined as

$$
D_{KL}(\mathbf{V}|\mathbf{W}\mathbf{H}) = \sum_{k,l}\left(\mathbf{V}_{k,l}\log\frac{\mathbf{V}_{k,l}}{(\mathbf{W}\mathbf{H})_{k,l}} - \mathbf{V}_{k,l} + (\mathbf{W}\mathbf{H})_{k,l}\right) \tag{8}
$$

which is a common choice for audio applications.

3.2. NMF on weakly labeled data

Let us consider the task of detection of rare sound events. Let $y \in \{0, 1\}$ be a weak label denoting absence or presence of the target sound, $\mathbf{V}^0 = \mathbf{V}_1^0, \cdots, \mathbf{V}_{M_0}^0$ is a set of M_0 training examples with absence of the target sound and $\mathbf{V}^1 = \mathbf{V}_1^1, \cdots, \mathbf{V}_{M_1}^1$ is a set of M_1 training examples with the presence of the target sound. As the data is weakly labeled, examples containing the target sound most probably also contain noise and other sounds. Therefore, we assume that to reconstruct well the target sound training examples ($\mathbf{V}^1$) we also need elements from dictionaries extracted from background sounds examples ($\mathbf{V}^0$). At the same time, we do not expect elements of the dictionary atoms of target sounds to be used for reconstructing $\mathbf{V}^0$. We impose this constraint in the training phase by applying a binary mask to the activation matrix as follows:

$$
\begin{aligned}
\mathbf{V} = [\mathbf{V}_0, \mathbf{V}_1] &\approx [\mathbf{W}_0, \mathbf{W}_1] \left(\begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \odot \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} \\ \mathbf{H}_{10} & \mathbf{H}_{11} \end{bmatrix}\right) \\
&= [\mathbf{W}_0, \mathbf{W}_1] \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{01} \\ \mathbf{0} & \mathbf{H}_{11} \end{bmatrix} \\
&= [\mathbf{W}_0\mathbf{H}_{00}, \mathbf{W}_0\mathbf{H}_{01} + \mathbf{W}_1\mathbf{H}_{11}]
\end{aligned} \tag{9}
$$

where $\mathbf{W}^0 \in {R_+}^{F\times K^0}$, $\mathbf{W}^1 \in {R_+}^{F\times K^1}$ are "sound" and "background" dictionaries respectively, K^0 and K^1 are their corresponding ranks. $\mathbf{0}$ is a matrix of zeros with K^1 rows and the number of columns corresponding to the total size of M_0 background training data, while $\mathbf{1}$ denotes matrices of appropriate dimensions with all elements equal to 1. $\mathbf{H}^{00}$, $\mathbf{H}^{01}$, $\mathbf{H}^{10}$ and $\mathbf{H}^{11}$ are parts of the activation matrix of suitable dimensions. We then further improve the separation of the dictionaries by adding an additional orthogonality regularizer, which minimizes the coherence between the dictionaries. Combining the constraint on the activation matrix and the orthogonality regularizer results in the following cost function to minimize:

$$
\begin{aligned}
&\min_{\mathbf{W}_0,\mathbf{W}_1,\mathbf{H}\geq 0} D_{KL}(\mathbf{V}|\mathbf{W}\mathbf{H}) + \lambda\|\mathbf{W}_1\mathbf{W}_0\|^2 \\
&= D_{KL}(\mathbf{V}_0|\mathbf{W}_0\mathbf{H}_{00}) + D_{KL}(\mathbf{V}_1|(\mathbf{W}_0\mathbf{H}_{01} + \mathbf{W}_1\mathbf{H}_{11})) \\
&+ \lambda\|\mathbf{W}_1\mathbf{W}_0\|^2.
\end{aligned} \tag{10}
$$

As $\|\mathbf{W}_1\mathbf{W}_0\|^2$ is convex in $\mathbf{W}_0$ and $\mathbf{W}_1$, we can minimize the cost function using the gradient descent. Then, following the derivations of Lee and Sung [22], we obtain the corresponding multiplicative update rules for $\mathbf{W}_0$ and $\mathbf{W}_1$:

$$
\begin{aligned}
\mathbf{W}_0 &\leftarrow \mathbf{W}_0 \odot \frac{\frac{\mathbf{V}_0\mathbf{H}_{00}}{\mathbf{W}_0\mathbf{H}_{00}} + \frac{\mathbf{V}_1\mathbf{H}_{01}}{\mathbf{W}_0\mathbf{H}_{01}+\mathbf{W}_1\mathbf{H}_{11}}}{\mathbf{1}\cdot\mathbf{H}_{00}{}^{+}\mathbf{1}\cdot\mathbf{H}_{01}{}^{+}\lambda\mathbf{W}_1\mathbf{W}_1\mathbf{W}_0} \\
\mathbf{W}_1 &\leftarrow \mathbf{W}_1 \odot \frac{\frac{\mathbf{V}_1\mathbf{H}_{11}}{\mathbf{W}_0\mathbf{H}_{01}+\mathbf{W}_1\mathbf{H}_{11}}}{\mathbf{1}\cdot\mathbf{H}_{11}{}^{+}\lambda\mathbf{W}_0\mathbf{W}_0\mathbf{W}_1}.
\end{aligned} \tag{11}
$$

As the regularizer does not influence the activation matrix $\mathbf{H}$, the update rule for $\mathbf{H}$ remains the same as in the original NMF problem formulation:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T \frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T \mathbf{1}}. \quad (12)$$

Final dictionaries $\mathbf{W}_0$ and $\mathbf{W}_1$ are then concatenated to form a dictionary $\mathbf{W}$ used for audio event detection, whereas the activation matrix $\mathbf{H}$ is discarded.

4. PROPOSED METHOD

For the final task of AED we combine the outputs of the two models. In principle, the salience model will detect onsets of any interesting events, regardless of whether they are the target in the task. Therefore, its output is combined with the NMF output, which can differentiate between target and non-target events. The full combined model is shown in Figure 1.

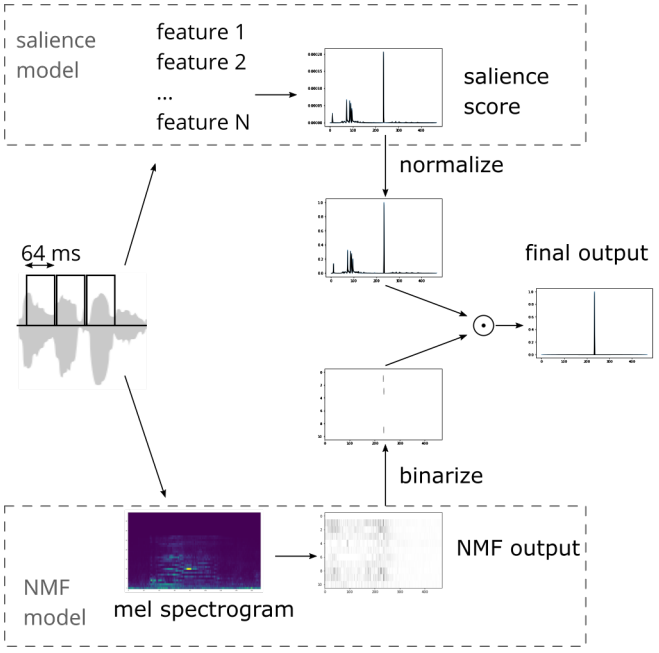

Fig. 1. General architecture of the proposed method. Frames of audio signal are analysed in parallel by two models. The salience model calculates salience score from N different features, while the NMF model uses mel spectrograms to produce an output matrix, which is binarized before being combined with the salience score to form the final output.

Firstly, we use the NMF method to extract dictionaries $\mathbf{W}_0$ and $\mathbf{W}_1$. In the event detection phase, a test sample is decomposed using the trained dictionaries as follows:

$$\mathbf{V}_{test} = \left[\mathbf{W}^0, \mathbf{W}^1\right] \begin{bmatrix} \mathbf{H}_0 \\ \mathbf{H}_1 \end{bmatrix}. \quad (13)$$

Finally, $\mathbf{H}_1$ is binarized using a threshold equal to 50% of the maximum value of the entire activation matrix ($\mathbf{H}_0$ and $\mathbf{H}_1$).

In parallel, salience score $s(n)$ per frame is computed for each test sample, forming a vector s, which is then normalised. In the final stage the salience score vector s is multiplied element-wise with the binary output of $\mathbf{H}_1$. The columns of the resulting matrix $\mathbf{O}$ that have at least one entry greater than 0 indicate the presence of an event.

$$\mathbf{O} = s \odot \mathbf{H}_1. \quad (14)$$

We post-process the annotation matrices by discarding events shorter than 100 ms and removing gaps shorter than 100 ms between the events. The shortest event in the training dataset was 240 ms long.

5. EXPERIMENTAL SETUP

5.1. Dataset

The proposed method is evaluated on rare event detection using only weakly labeled data from the audio recordings of the TUT Rare Sound Events 2017, which were provided for Task 2 of the DCASE2017 challenge [6]. Although the DCASE2018 Challenge also provided a weakly labeled dataset for AED, we choose the DCASE2017 dataset because it addresses a simple scenario - detecting one sound at a time - which is suitable for the current proof-of-concept study. The dataset consists of around 100 isolated sound examples for three target classes: gunshot, baby crying and glass breaking, together with background audio which is part of the TUT Acoustic Scenes 2016 dataset [23]. For training the NMF model we use weakly labeled 4 second mixtures. It is important to reiterate, that we do not know the timestamp of the event in the mix, just a binary label determining weather the mix contains the sound of interest. For testing, we use 500 mixes of -6dB, 0dB and 6dB SNR of the sounds and backgrounds not used in the training set. The testing mixtures were provided for the DCASE2017 challenge. Each testing mixture is 30 second long.

5.2. Parameters of the salience model

Six features are extracted using the pyAudioAnalysis library [24], with a 64 ms window: energy, energy entropy, spectral centroid, spectral rolloff, spectral entropy and zero-crossing rate. The weights w_{ii} and w_{ij} used in Eq. 6 were trained with a constrained logistic regression, where the binary output variable was presence of event in a file, predictor values were the mean s_i for each feature, and the weights were constrained to be positive. Recordings of 30 seconds are used for training.

5.3. Parameters of the NMF model

To build the NMF model, we resample the data to 16000 Hz. We extract mel-spectrograms with 40 components, using a window size and hop size of 64 ms. In order to model temporal dynamics we group 4 consecutive frames into 2D patches, a value that was chosen empirically. We train the system on audio chunks of 4 seconds and evaluate on recordings of 30 seconds. We set the number of positive and negative atoms as $K_0 = 20$, $K_1 = 10$, values chosen empirically.

5.4. Evaluation metrics

To evaluate the method we use event-based error rate (ER) and event-based F-score. An event is considered correctly detected using onset-only condition with a collar of 500 ms. The ER is calculated by adding the number of insertions and deletions for each class before dividing it by the total number of events. The F-score is based on the total amount of false negatives, true positives and false positives [25].

6. RESULTS AND DISCUSSION

Table 1 presents the results of the evaluation on the test set. For comparison, alongside the proposed method, we show the results for each of the NMF and salience models separately. In the salience model, an event was detected for every frame n in which $s(n) > 0.5 * \max(s)$.

Table 1. Evaluation results on gunshot, babycry and glassbreak detection. Error Rate (ER) and F-score (F1) are reported for the proposed method, NMF only and Salience model only. Lowest ER and highest accuracy for each target sound are shown in bold.

Event type	**Proposed**		NMF		Salience	
	ER	F1	ER	F1	ER	F1
Gunshot	**0.76**	**65%**	0.80	64 %	1.45	36%
Glass breaking	**1.07**	46%	1.23	41%	1.12	**54%**
Baby crying	**1.04**	36%	1.07	**37%**	1.71	32%

Adding the auditory salience model to the NMF detector improves its performance for gunshot and glass breaking events. For the baby crying event, it decreases the error rate, but does not improve the F-score, suggesting a low hit rate. The reason for this difference in performance for different event classes may be that the first two - gunshot and glass breaking - usually have sudden onsets, while the last one - baby crying - can start rather slowly. The salience model is designed to detect sudden changes in features, but will adapt to changes that are too slow. While this property makes it useful in some types of backgrounds (see below), it also means that it might not be suited for events which develop slowly, or might need a larger frame window for them.

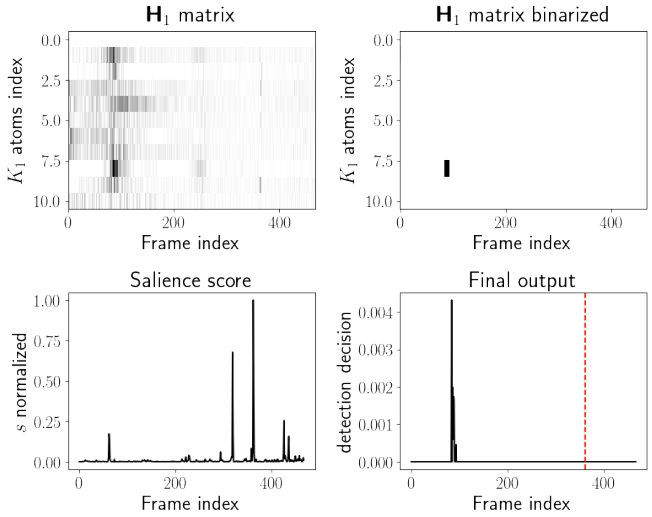

Fig. 2. Results for a gunshot event over a residential area background, with a loud car passing in the first half of the file. Top row: $\mathbf{H}_1$ matrix from the NMF model. Bottom left: salience model output s. Bottom right: final output of the model, from which an event is detected for any value larger than 0. Red dashed line shows the position of the target event. Even though it was correctly recognised by the salience model, the combined models do not detect it.

There were a number of cases where the salience model was able to detect an event when the NMF was not. This is also evident from the fact that the salience model outperforms both the NMF and the proposed method for the glass breaking event. One situation where the salience model presents an advantage is when the background noise significantly but slowly increases in level - e.g. a train passing (see Figure 2). Because a Kalman filter-based model is not sensitive to sudden feature changes, it is able to adapt to this background, and only flag a detection when changes in feature values correspond to new, 'surprising' events. It also seems to perform well in loud cafeteria-type backgrounds (see Figure 3).

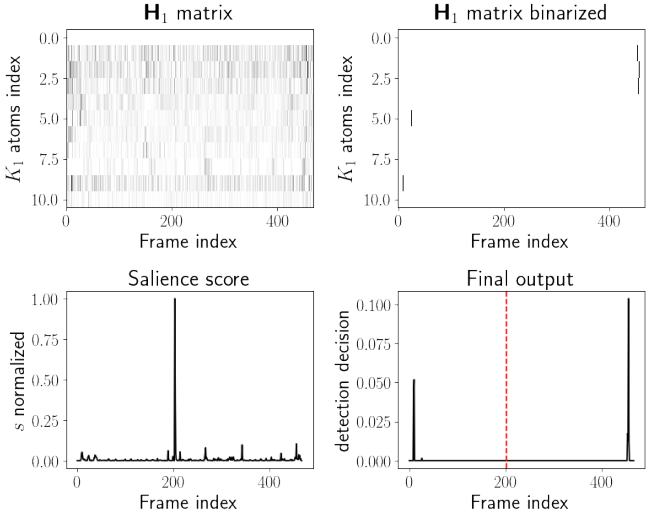

Fig. 3. Results for a gunshot event over a cafe/restaurant background. Top row: $\mathbf{H}_1$ matrix from the NMF model, before and after binarization. Bottom left: salience model output s, after normalization. Bottom right: final output of the model, from which an event is detected for any value larger than 0. Red dashed line shows the position of the target event, which was correctly recognised by the salience model, but not the NMF model.

For the combined model, however, binarization in the NMF output, followed by multiplication of outputs, does not allow for detection of any events not detected by NMF. With the method used to combine the two models, the main advantage of the salience model was in removing false positives. A different method, not based on a binary mask, might preserve more events detected by Kalman filters and increase hit rate.

7. CONCLUSIONS

We proposed a novel approach for AED using weakly labeled data: combining a traditional NMF based approach with a model inspired by human auditory attention. We showed that auditory salience can enhance traditional AED models. Specifically, a Kalman filter-based salience model provides promising results, as it seems less sensitive to changes in background sound level. However, due to the simplicity of the method for combining the two outputs, some of the salient sounds were removed from the final output by the NMF model. Therefore, in future we will explore more sophisticated methods for combining predictions from multiple models, for example a mixture of experts or other ensemble method. Moreover, it would be interesting to examine the combination of a salience model with a deep learning model, such as a Convolutional Recurrent Neural Network.

8. REFERENCES

[1] P. V. Hengel and J. Anemüller, "Audio event detection for in-home care," *Proc. of International Conference on Acoustics (NAG-DAGA)*, pp. 618–620, 2009.

[2] J. Kotus, K. Lopatka, and A. Czyzewski, "Detection and localization of selected acoustic events in acoustic field for smart surveillance applications," *Multimedia Tools and Applications*, vol. 68, no. 1, pp. 5–21, 2014.

[3] M. Bugalho, J. Portêlo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Proc. of the 10th International Conference of the International Speech Communication Association (Interspeech 2009)*, 2009, pp. 1151–1154.

[4] F. Meucci, L. Pierucci, E. Del Re, L. Lastrucci, and P. Desii, "A real-time siren detector to improve safety of guide in traffic environment," *Proc. of the 16th European Signal Processing Conference (EUSIPCO 2008)*, 2008.

[5] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 2, pp. 379–393, 2018.

[6] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 85–92.

[7] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, 2013.

[8] A. Diment, T. Heittola, and T. Virtanen, "Sound event detection for office live and office synthetic AASP challenge," in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2013)*, 2013, extended abstract.

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

[10] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 45–49.

[11] O. Dikmen and A. Mesaros, "Sound event detection using non-negative dictionaries learned from annotated overlapping events," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2013)*, 2013, pp. 5–8.

[12] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, 2016, pp. 6440–6444.

[13] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, 2018, pp. 121–125.

[14] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, vol. 15, no. 21, pp. 1943–1947, 2005.

[15] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011)*, 2011, pp. 1173–1179.

[16] E. M. Kaya and M. Elhilali, "Investigating bottom-up auditory attention," *Frontiers in Human Neuroscience*, vol. 8, p. 327, 2014.

[17] ——, "Modelling auditory attention," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1714, 2017.

[18] B. Schauerte and R. Stiefelhagen, "'Wow!' Bayesian surprise for salient acoustic event detection," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2013)*, 2013, pp. 6402–6406.

[19] M. Heilbron and M. Chait, "Great expectations: Is there evidence for predictive coding in auditory cortex?" *Neuroscience*, vol. 389, pp. 54–73, 2018.

[20] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[21] I. Sobieraj, L. Rencker, and M. D. Plumbley, "Orthogonality-regularized masked NMF for learning on weakly labeled audio data," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, 2018, pp. 2436–2440.

[22] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization." *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[23] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. of the 24th European Signal Processing Conference 2016 (EUSIPCO 2016)*, 2016.

[24] T. Giannakopoulos, "pyAudioAnalysis: An open-source Python library for audio signal analysis," *PLOS ONE*, vol. 10, no. 12, p. e0144610, 2015.

[25] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.