SOUND EVENT ENVELOPE ESTIMATION IN POLYPHONIC MIXTURES

Irene Martín-Morató[1], *Annamaria Mesaros*[2], *Toni Heittola*[2],
Tuomas Virtanen[2], *Maximo Cobos*[1], *Francesc J. Ferri*[1]

[1] Department of Computer Science, Universitat de València
[2] Laboratory of Signal Processing, Tampere University of Technology

ABSTRACT

Sound event detection is the task of identifying automatically the presence and temporal boundaries of sound events within an input audio stream. In the last years, deep learning methods have established themselves as the state-of-the-art approach for the task, using binary indicators during training to denote whether an event is active or inactive. However, such binary activity indicators do not fully describe the events, and estimating the envelope of the sounds could provide more precise modeling of their activity. This paper proposes to estimate the amplitude envelopes of target sound event classes in polyphonic mixtures. For training, we use the amplitude envelopes of the target sounds, calculated from mixture signals and, for comparison, from their isolated counterparts. The model is then used to perform envelope estimation and sound event detection. Results show that the envelope estimation allows good modeling of the sounds activity, with detection results comparable to current state-of-the art.

Index Terms— Sound event detection, Envelope estimation, Deep Neural Networks

1. INTRODUCTION

Sound event detection (SED) aims to detect presence of different sounds in an audio recording and provide a textual label, onset and offset times for each [1]. In real-life environments, where different sound events may overlap, an ideal SED system should be able to detect all such overlapping sounds. This case is referred to as polyphonic SED [2], and was studied in many different tasks: SED in synthetic audio [3], in real-life audio [3, 4, 5], rare SED [6, 7] and SED using weakly-labeled data [6, 8, 9]. In all these tasks, sound events had to be detected in polyphonic mixtures, with either overlapping target sounds, or significant background present. Most state of the art methods use deep learning, with convolutional and recurrent neural networks being the most prominent [4, 5, 7, 8].

This work has received funding from the European Research Council under the ERC Grant Agreement 637422 EVERYSOUND and from the Spanish government through grants TIN2014-59641-C2-1-P, TIN2014-54728-REDC, BIA2016-76957-C3-1-R, FPU14/06329.

The common representation of sound events in current systems is in the form of binary activity indicators for individual sound instances. However, this is a very rough approximation of the natural activity patterns of sounds in real-life. Often sounds have different non-binary activity patterns, for example moving sources such as car passing by or vehicle sirens exhibiting a fade-in/fade-out effect, or variations that are not accurately explained by the binary activity, such as footstep sounds on different surfaces.

This paper proposes use of non-binary activity indicators to characterize the temporal activity of sound events: instead of estimating a point when a sound event becomes active/inactive, we propose estimating its amplitude envelope. Other works using energy envelope information exist, such as [10] where the envelope is used to extract the significant parts of the sound before performing classification but, to the best of our knowledge, there are no published studies targeting envelope estimation. The use of values other than 0 and 1 as targets for the network in training, changes the setup from frame-based classification into regression, which in turn changes the optimization function in the training procedure to a regression appropriate one. The estimated envelopes are evaluated by comparing them with the envelopes calculated from the test data, using mean squared error. Additionally, the estimated envelopes can be transformed into binary activity indicators by setting a threshold and mapping the values above and below into 0 or 1 accordingly; this output is then evaluated against the reference annotations using F1-score and error rate.

The paper is organized as follows: Section 2 describes the approach for envelope estimation, Section 3 describes the methods and evaluation of the system, while Section 4 presents the dataset used in the experiments, the experimental results and discussion. Finally, Section 5 presents conclusions and future work.

2. ENVELOPE ESTIMATION

In real-life scenarios, the input acoustic signal to be analyzed is usually a polyphonic mixture of target sound events. These mixtures commonly contain background noise and a number of overlapping events from different classes. Identifying the

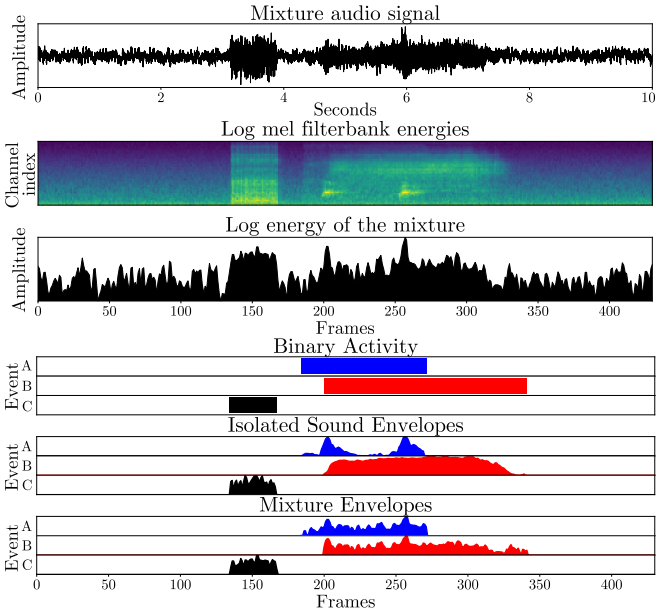

Fig. 1: The process of obtaining envelopes for the isolated sounds and the mixtures based on the binary activity indicators.

presence of sound events within the mixtures with binary indicators is sometimes difficult due to the variability of real-life sounds and the high level of polyphony. For example, a situation with two overlapping sounds produced by moving sources (e.g. car passing by) that have first a gradual increase of energy as they come nearer to the observer and then a gradual decrease as they move away, is hard to describe using only binary indicators. Instead, we can try to estimate an accurate representation of the mixture signal, identifying the progressive presence or absence of the events present in it. This representation can be a distribution of the acoustic signal in the continuous domain offering more precise information about the acoustic events. With this continuous range it is possible to mark the gradual presence of the target sound activity with a wider range of values, not only 0s and 1s. For obtaining such a representation, we propose to estimate the amplitude envelope of the acoustic signal. We represent the envelope by calculating the logarithm of the energy of the acoustic signal in the time domain. The use of the logarithm provides a smoother representation of the temporal evolution of the energy, leading to better envelope estimation results.

Learning of sound envelopes is based on training data, for which we obtain the envelope information as illustrated in Figure 1. The main assumption of the proposed method is that given an acoustic signal, if the target sound event is in the foreground, the energy of the signal within its temporal vicinity will reflect the activation of this sound. For extracting the envelope, we consider a mixture signal in which the target sounds have been annotated with binary activity indicators. We estimate the amplitude envelope of the energy of the signal and multiply it with the binary activity of each annotated sound instance, to obtain the activity information within the annotated segment; this is further normalized to obtain values between 0 and 1 for that sound event instance.

In order to investigate the effect of this approximation on the output of the system, we use synthetic mixtures to train and evaluate the proposed method. This allows us to access the precise envelopes by calculating them from the isolated sound instances, and comparing them with the envelopes calculated from the mixture signal. Figure 1 illustrates this comparison. For non-overlapping sounds, such as the sound labeled C in the figure, the difference in the resulting envelope is small, but for the sounds that overlap, the resulting shape can be dramatically different, as observed for sounds labeled A and B. Our hypothesis is, however, that envelopes obtained from mixtures can be successfully used for training.

3. METHODS AND EVALUATION

3.1. System design

The model architecture used in this work is a Convolutional Recurrent Neural Network (CRNN) based on the system proposed in [4] that ranked first for sound event detection in real-life audio in DCASE 2017 challenge. The first layers are CNN, each of them followed by batch normalization and max-pooling. The output of the CNN is fed to bi-directional gated recurrent units (GRU), which learn the temporal activity patterns. The last layers are time-distributed fully-connected (dense) layers. The output layer has sigmoid activation, so it can produce multi-label output. The input to the neural network consists of T consecutive time frames of mel-band energies N_{mbe}; the dimensions are $T = 431$ given by the length of the audio files and $N_{mbe} = 40$ number of mel-bands in the frequency range of $0 - 22500Hz$.

For training with the envelopes, the optimization loss function used is the mean squared error (MSE) instead of the usual binary cross-entropy used for training systems for classification. The best values for batch size and binarization threshold used to transform the regression output into detection are selected using the validation set. The values we find worked best are batch size of 32 for mixtures, 16 for the isolated events; for both cases the best binarization threshold was 0.25. Training was performed using Adam optimizer [11] with a learning rate of 0.001.

For comparison, we use the same system trained for detection, as in the original work. For the detection case, the architecture and features stay the same, but the targets are binary. The optimization function used during training is binary cross-entropy, and the output values are thresholded ($threshold = 0.5$) to obtain the final binary decision. Training was performed for 500 epochs with a batch size of 32.

The optimal binarization threshold for the envelope estimation is smaller than the one used with binary labels, because we have continuous values that represent the in-

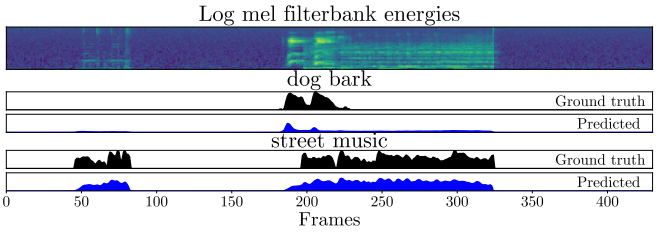

Fig. 2: Envelopes estimated by the system trained with isolated sound envelopes.

cremental presence of an event, and therefore the model is expected to predict sound presence using smaller values. Figure 2 presents one example of ground truth and predicted output in which the training envelopes are calculated using the isolated sounds. It can be seen that in some regions the event presence is marked by low values.

3.2. Evaluation

We evaluate the system output both from the envelope estimation and SED perspectives. Because the envelope estimation is a regression problem, we evaluate its output using the MSE between the system output and the data points. In order to separate the system behavior between the active and inactive regions of the target sounds, we calculate MSE separately for these regions, according to the reference annotations. Furthermore, because MSE is difficult to interpret due to arbitrariness of its scale, we calculate SNR of the estimated envelopes by dividing the energy of the reference envelopes ($Energy_{ref}$) to the squared error:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{Energy_{ref}}{Error} \right), \qquad (1)$$

where $Error = \sum_{n=1}^{T} (ref[n] - pred[n])^2$, calculates the difference between the reference (ref) and predicted ($pred$) envelopes along time T.

To evaluate SED, we transform the regression output into binary activity indicators using a threshold: all values above the selected threshold are considered 1, and all below are considered 0. This output is further processed by imposing a gap of at least 0.1 s between active blocks in order to consider them as different event instances, and imposing a minimum sound event length of 0.1 s. The final output is then evaluated using segment-based error rate ER and F1-score in 1 s segments [2].

4. EXPERIMENTAL RESULTS

4.1. Audio data

For this study we use the URBAN-SED dataset created using Scaper [12]. The dataset contains mixtures of urban sounds from the UrbanSound8k dataset [13] which is distributed into 10 stratified folds and contains 10 different classes: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren and street music*. The data is divided into training (6000 soundscapes from folds 1-6), validation (2000 soundscapes from folds 7-8) and test data (2000 soundscapes from folds 9-10). The mixtures are generated by selecting the same background Brownian noise for all the files. For generating a high variable set of mixtures, a collection of parameters are used for modifying the sound events before adding them to the mixture (e.g. start time , duration); for details, please refer to [12].

Given the synthetic generation of the dataset, the annotations are guaranteed to be correct and complete, compared to the uncertainty of manually annotated datasets. The dataset also contains sound events such as dog barking and children playing that have fluctuating envelopes. In addition, this dataset allows us to verify our hypothesis that the events in the foreground can be represented using the mixture signal energy, by comparing the use of envelopes obtained from the original isolated sounds and from the mixture signal.

Event class	MSE	SNR [dB]
air conditioner	0.190	2,658
car horn	0.181	3,459
children playing	0.152	3,198
dog bark	0.168	2,609
drilling	0.172	3,328
engine idling	0.148	3,917
gun shot	0.136	2,728
jackhammer	0.077	6,745
siren	0.129	4,187
street music	0.138	3,799

Table 1: Mean squared error of regression output and Signal to Noise Ratio (SNR) for active regions of the target sounds; training using mixture envelopes.

4.2. Envelope estimation results

Class-wise results for the envelope estimation regression problem are presented in Table 1, which include MSE and SNR from the active regions or the target sounds. Based on MSE, we can conclude that the system performs in quite consistent manner, with MSE being within close range for all classes. However, it is hard to assess what is a good MSE value. Based on the SNR, we can interpret the scale of the errors with respect to the reference signal: the jackhammer and siren class are estimated best, while for air conditioning, dog bark and gun shot the estimation has the highest error.

We also calculated MSE in the inactive regions of each sound class, and obtained for all classes value in the range of 0.002-0.009–meaning that the system correctly predicts values close to zero in the inactive regions of all sound event classes. If we transform the regression to detection (as eval-

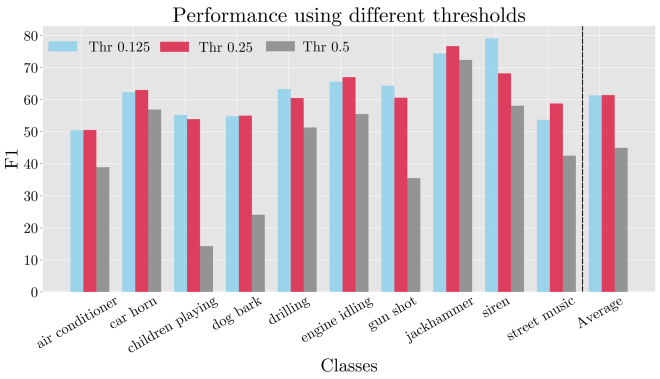

Fig. 3: F1-score in 1 s segments for different binarization thresholds; training using envelopes from mixtures

uated in the next subsection), a close inspection of error rates produced by the system shows that the insertion rate is very small for all classes, with the vast majority of the errors produced being deletions. This explains why in the inactive regions the regression output is mostly correct.

4.3. Sound event detection results

For comparison with published work, we choose segment based ER and F1 score in 1 s segments [2], and present class-wise results (macro-averaging) as well as instance-wise (micro-average). Figure 3 presents the class-wise detection results with different binarization thresholds; the average values for thresholds 0.125 and 0.25 are very close (61.31 vs 61.42), but based on the validation data, the threshold of 0.25 is selected as the one leading to best ER and F1-score.

Table 2 presents the performance comparison between detection with binary activity and detection through envelope estimation, with the training envelopes based on the isolated sounds and on the mixture audio. The system using binary information was not optimized further, therefore had the 0.5 threshold; we compare it with the best result obtained by the envelope estimation system, which is for a 0.25 threshold. Results in Table 2 show that our reference system trained using binary activity indicators has a higher performance than the system described and analyzed in [12]. We therefore consider that our reference system is a reasonably good representation of current state-of-the art performance

Regression-based detection has a slightly lower average performance, with the system trained with envelopes calculated from mixtures having lowest performance, but still few percent units higher than [12]. Class-wise performance is very similar for sound events that have a more stationary nature, like air conditioning, engine idling, siren, while for sounds that have a more dynamic structure, performance of detection using envelope estimation is smaller. The largest performance gap of 10% is for gunshot, probably because the energy envelope has only few values that provide information, while the binary activities give more weight to the "tail" of the sound. Since very short events in the regression output are filtered out by the postprocessing, it may also be the case that some detected very short gunshot events are discarded.

Event class	binary	isolated env.	mixture env.
air conditioner	48.3	49.2	50.5
car horn	66.0	66.9	63.0
children playing	56.9	56.7	53.9
dog bark	60.3	59.6	55.0
drilling	66.3	63.0	60.5
engine idling	68.2	67.0	67.0
gun shot	71.5	60.7	60.6
jackhammer	78.3	78.6	76.7
siren	69.9	69.0	68.2
street music	59.7	60.5	58.8
average	64.3	63.1	61.4

Table 2: F1-score in 1 s segments for different approaches to detection; estimated envelopes binarize with 0.25 threshold

System training	F1	ER
binary activity	64.7	0.48
envelope from isolated examples	63.6	0.49
envelope from mixture signal	61.8	0.52

Table 3: F1-score and error rate calculated using micro-averaging (1 s segment-based)

For completeness, we evaluate the detection results using error rate and F1-score as used in DCASE Challenge. The difference to the evaluation in Table 2 is the overall accumulation of counts before metric calculation (micro-average) instead of the class-wise metrics. However, the difference is rather small because the system performance is consistent between classes, and the dataset is rather balanced. The presented results show that estimating the sound envelopes provides SED results comparable with state-of-the-art performance.

5. CONCLUSIONS AND FUTURE WORK

We have presented an approach for estimating the envelope of sound events in polyphonic mixtures. Envelope estimation results evaluated by MSE and SNR show the effectiveness of the method. In addition, the envelopes as activity descriptors were transformed into binary activity indicators for estimation of SED capability of the method. The proposed approach has comparable performance to a state-of-the-art system trained using binary labels, therefore we can conclude that estimation of envelopes can provide satisfying performance in SED. To validate the current conclusions, future work will target application of the method for real-life recordings, where the training envelopes are not available from isolated examples, but can be calculated only based on the mixture and corresponding annotations.

6. REFERENCES

[1] T. Heittola, E. Çakır, and T. Virtanen, *The Machine Learning Approach for Analysis of Sound Scenes and Events*, pp. 13–40, Springer International Publishing, Cham, 2018.

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

[3] 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 Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[4] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 771–775.

[5] I-Y. Jeong, S. Lee, Y. Han, and K. Lee, "Audio event detection using multiple-input convolutional neural network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 51–54.

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

[7] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 80–84.

[8] D. Lee, S. Lee, Y. Han, and K. Lee, "Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 74–79.

[9] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, July 2018.

[10] I. Martín-Morató, M. Cobos, and F. J. Ferri, "Adaptive mid-term representations for robust audio event classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2381–2392, Dec 2018.

[11] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[12] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017, pp. 344–348.

[13] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.