ROBUST SOUND EVENT DETECTION THROUGH NOISE ESTIMATION AND SOURCE SEPARATION USING NMF

Qing Zhou, Zuren Feng

Xi'an Jiaotong University
School of Electronic and Information Engineering
No. 28 Xianning West Road, Xi'an, Shaanxi 710049, P. R. China
belief2012@stu.xjtu.edu.cn, fzr9910@mail.xjtu.edu.cn

ABSTRACT

This paper addresses the problem of sound event detection under non-stationary noises and various real-world acoustic scenes. An effective noise reduction strategy is proposed in this paper which can automatically adapt to background variations. The proposed method is based on supervised non-negative matrix factorization (NMF) for separating target events from noise. The event dictionary is trained offline using the training data of the target event class while the noise dictionary is learned online from the input signal by sparse and low-rank decomposition. Incorporating the estimated noise bases, this method can produce accurate source separation results by reducing noise residue and signal distortion of the reconstructed event spectrogram. Experimental results on DCASE 2017 task 2 dataset show that the proposed method outperforms the baseline system based on multi-layer perceptron classifiers and also another NMF-based method which employs a semi-supervised strategy for noise reduction.

Index Terms— Sound event detection, non-negative matrix factorization, sparse and low-rank decomposition, source separation

1. INTRODUCTION

Sound events such as gunshots, screams, glass breaks, etc. are often associated with hazardous situations. Automatic detection and monitoring of these sound events can be very useful for security reasons. A key problem in sound event detection is the presence of the highly non-stationary and time-varying background noise in realistic applications [1]. Most existing methods using a well-trained classifier on environment-specific training data is designed for particular situations and is thus unable to handle unseen noise. In addition, even if it is possible to train a classifier with an enormous amount of data involving different types of sounds in different environments, it enables the flexibility in dealing with different noises but at the expense of sacrificing performance at specific environments [2]. The goal of this paper is to develop robust detection methods which can automatically adapt to background variations for practical applications.

Techniques of non-negative matrix factorization (NMF) [3] have been extensively studied and successfully applied in speech enhancement for separating speech from noise [4-6]. Recently, many sound event detection systems using NMF have been published with promising results [7-11]. NMF models the spectrogram of a sound signal with a dictionary of spectral bases and a corresponding activation matrix. Since the activations may vary along time, this model can describe non-stationary signals to some extent. The key strategy for NMF to detect target events from noise is to express noise and target events by different sets of bases. The input noisy signal is first decomposed by this combined dictionary and then the target events can be reconstructed by only using the event components.

In sound event detection task, samples of the target event class are usually available and an event dictionary can be pre-trained and kept fixed during test. Strategies differ when dealing with noise. If a noise dictionary is also pre-trained and used in the decomposition, it is the supervised case. In contrast, in the semi-supervised case the noise dictionary is unknown and needs to be updated concurrently during test. For example, Gemmeke et al. [7] extracted bases for both the target event and the background noise and kept the dictionaries fixed during test. But this method can only be applied in simple and fixed noise conditions. To better handle unseen noise, Komatsu et al. [8] adopted the semi-supervised NMF strategy. A noise dictionary was introduced and learned during test with the aim of modeling unknown spectra which were not included in the training data. Their method can adapt to different noises but is not suitable for handling non-stationary noises. Since it lacks control over the noise bases, this method may not obtain accurate separation results, especially when there are many interference sound signals in the background which exhibit similar spectral profiles as the target event.

In order to enhance system robustness and reduce background interference, this paper proposes to first estimate a noise dictionary from the input test signal and then conducts the supervised source separation procedure. Noise dictionary learning is accomplished by the technique of sparse and low-rank decomposition [4, 12, 13] which has been proved useful in foreground/background separation. The underlying idea is to decompose a matrix into the summation of a low-rank matrix and a sparse matrix. This model also applies to sound event detection, that is, the foreground events are sparse due to its rare occurrence while the background noise is usually more stable and also less spectrally diverse than the foreground events and thus can be modeled by the low-rank part [13]. The low-rank part is further expressed by a linear combination of a limited number of bases (referring to the noise dictionary) upon NMF. Guided by the additional knowledge of noise bases, the event spectrogram can be better reconstructed via supervised NMF with less noise residue and signal distortion. The proposed method shows its effectiveness of adaptive noise reduction and achieves better performance compared to the semi-supervised method especially when dealing with similar background interference.

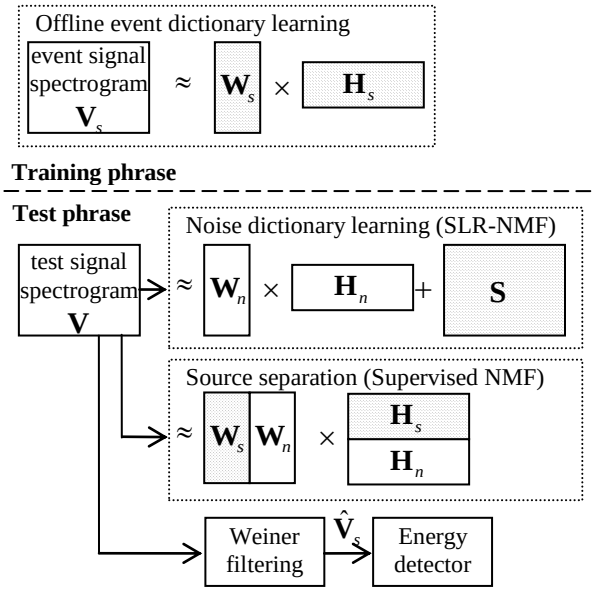

Figure 1: Framework of the proposed method

2. PROPOSED METHOD

The framework of the proposed method is presented in Fig. 1. In the training phase, an event dictionary for the target event class is trained by unsupervised NMF using the training set of clean event samples. The test phrase mainly consists of three steps: noise dictionary learning, supervised source separation, and event detection.

First, a noise dictionary is estimated from the input test signal via unsupervised sparse and low-rank non-negative matrix factorization (called SLR-NMF for short in this paper). Then, combining the estimated noise dictionary with the pre-trained event dictionary, the input signal is decomposed again by supervised NMF for source separation. Thus the event spectrogram can be reconstructed by only using the event components. Finally, the estimated event spectrogram is further smoothed and then processed by an energy detector to generate the final onset/offset results.

2.1. Noise dictionary learning by SLR-NMF

Sparse and low-rank decomposition represents a matrix as the summation of a low-rank matrix and a sparse matrix. The interpretation with respect to the low-rank part and the sparse part differs according to specific applications. For sound event detection, the foreground events rarely happen and occupy very limited entries of the input matrix and thus can be well expressed by the sparse part. In contrast, the background noise is usually more stable and also less spectrally diverse than the foreground events and thus can be modeled by the low-rank part.

Let $\mathbf{V} \in \mathbb{R}_+^{N\times T}$ denote the spectrogram of the input noisy signal where N is the number of frequency bins and T is the number of time frames. The model of SLR-NMF is given by

$$\mathbf{V} \approx \mathbf{W}_n\mathbf{H}_n + \mathbf{S} \tag{1}$$

in which $\mathbf{S} \in \mathbb{R}_+^{N\times T}$ is the sparse part representing the foreground events, and the low-rank part dedicated to the background noise is further represented upon NMF as the product of a noise dictionary $\mathbf{W}_n \in \mathbb{R}_+^{N\times R_n}$ and an activation matrix $\mathbf{H}_n \in \mathbb{R}_+^{R_n\times T}$. R_n is the number of noise bases satisfying $R_n < \min\{N,T\}$ and the subscript *n* refers to "noise".

The following optimization problem is constructed to solve the decomposition in (1):

$$\min_{\mathbf{W}_n,\mathbf{H}_n,\mathbf{S}} D\left(\mathbf{V} \mid \mathbf{W}_n\mathbf{H}_n + \mathbf{S}\right) + \lambda\|\mathbf{S}\|_1 \tag{2}$$

in which the first term is the Kullback-Leibler (KL) divergence [14] between the input matrix and its approximation, and the second term is a sparsity constraint on **S** which is measured by its L_1-norm. λ controls the weight of the sparsity constraint in the cost function and its selection will be discussed later in Section 3.

The multiplicative update rules for (2) are given as follows:

$$\mathbf{W}_n \leftarrow \mathbf{W}_n \odot \left(\frac{\mathbf{V}}{\mathbf{W}_n\mathbf{H}_n + \mathbf{S}}\mathbf{H}_n^{\mathrm{T}}\right) \Big/ \left(\mathbf{1}\mathbf{H}_n^{\mathrm{T}}\right) \tag{3}$$

$$\mathbf{H}_n \leftarrow \mathbf{H}_n \odot \left(\mathbf{W}_n^{\mathrm{T}}\frac{\mathbf{V}}{\mathbf{W}_n\mathbf{H}_n + \mathbf{S}}\right) \Big/ \left(\mathbf{W}_n^{\mathrm{T}}\mathbf{1}\right) \tag{4}$$

$$\mathbf{S} \leftarrow \mathbf{S} \odot \left(\frac{\mathbf{V}}{\mathbf{W}_n\mathbf{H}_n + \mathbf{S}}\right) \Big/ \left(\mathbf{1} + \lambda\right) \tag{5}$$

where **1** is an all-1 matrix with the same dimension as **V** and the superscript T means the transposition of a matrix. $\mathbf{A}\odot\mathbf{B}$ and $\mathbf{A}/\mathbf{B}$ refer to the element-wise multiplication and division, respectively. Hence, by solving (2) we obtain an estimate of the noise dictionary which directly describes the surrounding background of the current input signal.

2.2. Source separation by supervised NMF

Incorporating the noise dictionary $\mathbf{W}_n$ learned in Section 2.1 and the pre-trained event dictionary, the input test signal can be decomposed into the noise part and the event part by NMF in a supervised way as follows:

$$\mathbf{V} \approx \mathbf{W}_s\mathbf{H}_s + \mathbf{W}_n\mathbf{H}_n \tag{6}$$

in which $\mathbf{W}_s \in \mathbb{R}_+^{N\times R_s}$ with R_s bases refers to the event dictionary and the subscript *s* refers to "event signal". $\mathbf{W}_s\mathbf{H}_s$ and $\mathbf{W}_n\mathbf{H}_n$ are the estimated event and noise spectrograms, respectively. Here we conduct a second separation to the input noisy signal in a way that is different from the totally unsupervised decomposition in Section 2.1. This step can obtain more reliable separation results since both the prior knowledge of the noise and the target event class to be detected is utilized.

Supervised NMF in (6) is solved by minimizing the KL divergence between the input matrix and its reconstruction. The corresponding optimization problem is expressed as

$$\min_{\mathbf{H}_s,\mathbf{H}_n} D\left(\mathbf{V} \mid \mathbf{W}_s\mathbf{H}_s + \mathbf{W}_n\mathbf{H}_n\right) \tag{7}$$

Update rules for the activation matrices $\mathbf{H}_s$ and $\mathbf{H}_n$ in (7) are given by

$$\mathbf{H}_s \leftarrow \mathbf{H}_s \odot \left(\mathbf{W}_s^{\mathrm{T}} \frac{\mathbf{V}}{\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_n\mathbf{H}_n} \right) \Big/ \left(\mathbf{W}_s^{\mathrm{T}} \mathbf{1} \right) \quad (8)$$

$$\mathbf{H}_n \leftarrow \mathbf{H}_n \odot \left(\mathbf{W}_n^{\mathrm{T}} \frac{\mathbf{V}}{\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_n\mathbf{H}_n} \right) \Big/ \left(\mathbf{W}_n^{\mathrm{T}} \mathbf{1} \right) \quad (9)$$

2.3. Event detection

Like what is done in speech processing [4-5], Wiener filtering is conducted on the input spectrogram to get the final estimate of the event spectrogram, that is,

$$\hat{\mathbf{V}}_s = \mathbf{V} \odot \frac{\mathbf{W}_s\mathbf{H}_s}{\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_n\mathbf{H}_n} \quad (10)$$

For event detection, an energy detector is applied to $\hat{\mathbf{V}}_s$ by measuring the accumulation of energies of all frequency bins per frame. High energy values exceeding a threshold in a number of successive frames indicate the presence of a target event.

3. EXPERIMENTS

The proposed method is evaluated on DCASE 2017 task 2 dataset. This task focuses on the detection of three types of rare sound events (baby cry, glass break, and gunshot) in artificially created mixtures. The background audio material is from 15 different audio scenes including home, park, metro station, etc., making it a very complex detection scenario. In our detection algorithm, only isolated clean sound examples for target event classes are used for training and an event dictionary for each target event class is trained separately. It should be pointed out that no background audio is used for training. The development test set which contains 500 mixture audio examples for each target class is used for evaluation. Different SNR levels are tested including 6 dB, 0 dB, and –6 dB. All audio files are resampled to a sampling rate of 44100 Hz. Magnitude spectrogram via the short-time Fourier transform (STFT) is extracted as audio features using a sliding window with a frame length of 40 ms and 50% overlap.

Metrics used in the challenge are event-based error rate (ER) and event-based F-score (see [15] for details). An event is considered correctly detected using onset-only condition with a collar of 500 ms.

3.1. Parameter settings

Major parameters in the proposed algorithm are R_s, R_n, and λ. We use the training mixtures in the development dataset for tuning the parameters, which is disjoint from the test set. The search range for each parameter is empirically determined, that is, $R_s, R_n \in \{16, 32, 48, 64\}$, $\lambda \in \{0.05, 0.1, 0.2, 0.5, 0.8, 1\}$.

The numbers of event bases and noise bases are very important parameters. For qualitative analysis, using a sufficient number of bases is preferable to model audio sources precisely. However, using too many bases may degrade the performance since it would easily lead to the mixing problem, that is, noise may be wrongly described by the event bases or in the reverse way. After a grid search over the pre-defined range, we found that R_s=32 and R_n=32 are good choices which guarantee excellent performance and also a satisfactory computational load.

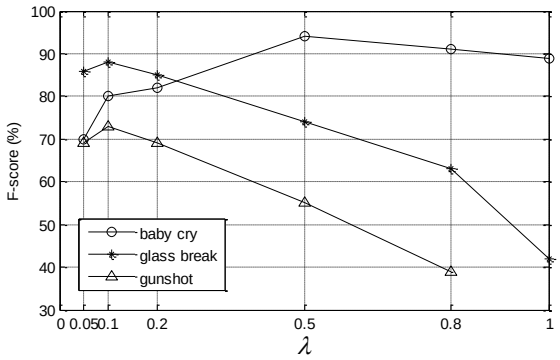

Figure 2: Detection results on the training mixtures under different values of the sparsity parameter λ. Best results are achieved at 0.1 for glass breaks and gunshots while 0.5 for the baby cry class.

The sparsity parameter λ used in the noise dictionary learning step has a significant effect on performance. It controls the strength of the sparsity constraint on the foreground event part and thus determines a trade-off between noise reduction and signal distortion. A larger λ means a sparser foreground estimate and a more sufficient noise estimate but at the expense of including some foreground event components. Since the goal of this decomposition is to learn a noise dictionary, it is better not to retain too many foreground event components within the noise part. So λ should be a relatively small value. According to the F-score results in Fig. 2, best results for glass breaks and gunshots are achieved around λ=0.1. The performance degrades a bit under 0.1 and a larger λ also yields poor results. However, the case for the baby cry class differs and it turns out that a large λ produces better results. This may be attributed to the considerable difference between the baby cry spectrum and the noise spectrum which enables a tolerance of the event residue within the noise part. Hence we set λ=0.5 for the baby cry class and λ=0.1 for glass breaks and gunshots in experiments.

For post-processing, the energy sequence computed from the estimated event spectrogram is further smoothed by a moving average filter. The duration of the filter is set to be a little shorter than the minimum length of the target event class as in [7].Very short detected events are also removed. Additionally, in order to obtain more accurate onset/offset results, a double-thresholding strategy is adopted. In other words, events are first detected using a large threshold, and then a small threshold is used to search within a small range before and after the duration of the detected event for the final onset and offset. This strategy is necessary and found to be useful especially for the baby cry class in experiments. Because there might be several phrases or short pauses within a baby cry event, the detector using one threshold may only locate the phrase with higher energies thus resulting in inaccurate onset or offset. The number of iterations in all the NMF algorithms is set to be 200 which is sufficient for convergence in our experiments.

Table 1: Evaluation results on the development test set for each target event class

Event class	Proposed method		Semi-supervised NMF	
	ER	F-score	ER	F-score
Baby cry	0.20	89.9%	0.27	83.5%
Glass break	0.22	89.2%	0.31	80.1%
Gunshot	0.42	78.4%	0.57	65.8%
Average	0.28	85.8%	0.39	76.5%

3.2. Results and discussions

The detection results for each target event class on the development test set are presented in Table 1. Compared with the baseline results provided by the challenge [16-17], our method achieves better performance. The average ER and F-score of the proposed method are 0.28 and 85.8% compared to 0.53 and 72.7% of the baseline results. The baseline system employs a multi-layer perceptron (MLP) classifier for each target class and needs to be trained on mixture audio examples of different scenes and different SNR levels. In contrast, our method only uses isolated event examples for training and adopts an adaptive noise reduction strategy, which is an advantage over the baseline system.

We also compared the proposed method with the semi-supervised NMF method which was applied in [8]. It should be explained that the fully supervised NMF method which uses an offline-trained noise dictionary is not considered in this paper since it is not suitable for DCASE2017 task 2 which covers various noise scenes. Training a noise dictionary on such a large amount of background data involving different sounds in different scenes is troublesome. Moreover, it may lead to a large overlap between the feature space spanned by the noise bases and that of the target event, which can greatly degrade performance.

It can be observed from Table 1 that our method outperforms the semi-supervised one and the detection results are improved for all three event classes. The proposed method estimates a noise dictionary from the current input signal which directly models the surrounding background and thus can obtain accurate separation results in the supervised NMF step. However, the semi-supervised method lacks control over the noise bases and may wrongly decompose noise into the event part or conversely. So it can easily lead to noise residue or signal distortion within the estimated event spectrogram. The situation gets worse especially when encountering analogous background interference.

The comparison of the two methods is illustrated by two test examples as shown in Fig. 3. In the baby cry detection example, there existed strong noise interference in the test signal. Our method had a fairly good effect of noise reduction and obtained excellent detection results. As for the semi-supervised method, the baby cry event was well captured by the event bases but part of the noise were also describe by the event bases, which led to harmful confusion in detection. In the second example of gunshot detection, the background noise in the test signal was highly non-stationary and contained sounds whose spectral profile was very similar to that of a gunshot event. Our method demonstrated its ability of tackling similar background interference and correctly located the gunshot event. However, it can be seen from the estimated event part of the semi-supervised method that many important event components were lost.

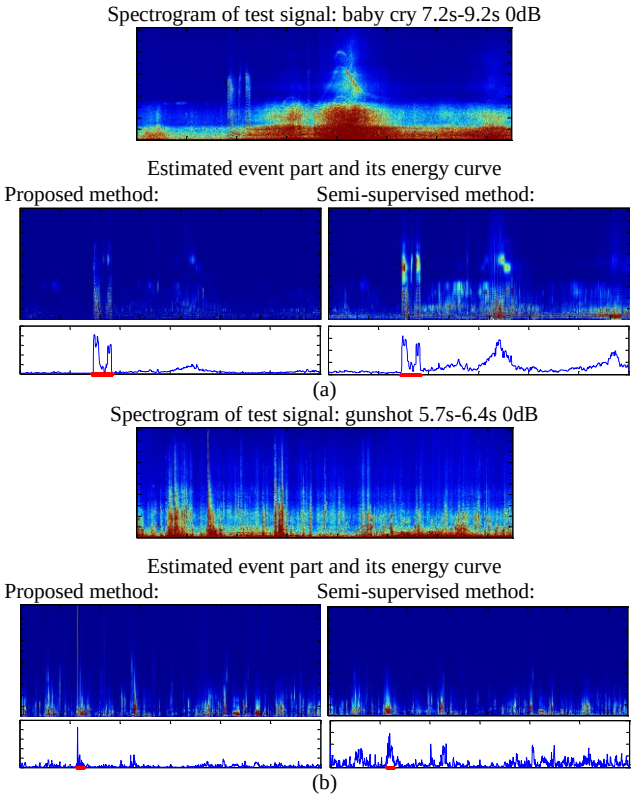

Figure 3: Detection examples of (a) a baby cry event and (b) a gunshot event

4. CONCLUSIONS

This paper presents a sound event detection method based on source separation by supervised NMF. In order to deal with non-stationary noises and background variations, this paper proposes to estimate a noise dictionary online from the input test signal using the technique of sparse and low-rank decomposition. Because the estimated noise dictionary can exactly describe the surrounding background, the succeeding supervised source separation via NMF can provide accurate separation of target events and noise. Experimental results demonstrate the noise reduction ability of the proposed method when dealing with non-stationary noises and similar background interference. The proposed method achieves better results than the baseline system based on MLP and also outperforms another NMF-based method which employs a semi-supervised strategy for noise reduction. Note that the proposed work needs a relatively long signal to learn a noise dictionary. It is not suitable for real-time applications. Future work will be dedicated to develop real-time noise dictionary learning techniques.

5. REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, V. Murino, "Audio surveillance: a systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 52:1–52:46, 2016.

[2] A. Rabaoui, Z. Lachiri, and N. Ellouze, "Using HMM-based classifier adapted to background noises with improved sounds features for audio surveillance application," *Int. J. Signal Process.*, vol. 5, no. 1, pp. 46–55, 2009.

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

[4] M. Sun, Y. Li, J. F. Gemmeke, and X. W. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with K-L divergence," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[5] M. N. Schmidt, J. Larsen, and K. Lyngby, "Wind noise reduction using non-negative sparse coding," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2007, pp. 431–436.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[7] J. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme, "An exemplar-based NMF approach to audioevent detection," in *Proc. WASPAA*, 2013, pp.1–4.

[8] T. Komatsu, T. Toizumi, R. Kondo, and S. Yuzo, "Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionary," in *Proc. DCASE2016 Workshop*, 2016, pp. 45–49.

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

[10] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Proc. CHiME*, 2011, pp. 36–40.

[11] T. Heittola, A. Mesaros, T. Virtanen, and G. Moncef, "Supervised model training for overlapping sound events based on unsupervised source separation," in *Proc. ICASSP*, 2013, pp. 8677–8681.

[12] P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP*, 2012, pp. 57–60.

[13] Z. Chen and D. P. W. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *Proc. WASPAA*, 2013, pp.1–4.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *NIPS*, 2001, pp. 556–562.

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

[16] 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. DCASE2017 Workshop*, 2017, submitted.

[17] DCASE2017 Challenge, "IEEE ASSP Challenge of Detection and Classification of Acoustic Scenes and Events," http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/.