BLSTM-HMM HYBRID SYSTEM COMBINED WITH SOUND ACTIVITY DETECTION NETWORK FOR POLYPHONIC SOUND EVENT DETECTION

Tomoki Hayashi[1], *Shinji Watanabe*[2], *Tomoki Toda*[1], *Takaaki Hori*[2], *Jonathan Le Roux*[2], *Kazuya Takeda*[1]

[1]Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan
[2]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA

hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp,
{takeda,tomoki}@is.nagoya-u.ac.jp, {watanabe,thori,leroux}@merl.com

ABSTRACT

This paper presents a new hybrid approach for polyphonic Sound Event Detection (SED) which incorporates a temporal structure modeling technique based on a hidden Markov model (HMM) with a frame-by-frame detection method based on a bidirectional long short-term memory (BLSTM) recurrent neural network (RNN). The proposed BLSTM-HMM hybrid system makes it possible to model sound event-dependent temporal structures and also to perform sequence-by-sequence detection without having to resort to thresholding such as in the conventional frame-by-frame methods. Furthermore, to effectively reduce insertion errors of sound events, which often occurs under noisy conditions, we additionally implement a binary mask post-processing using a sound activity detection (SAD) network to identify segments with any sound event activity. We conduct an experiment using the DCASE 2016 task 2 dataset to compare our proposed method with typical conventional methods, such as non-negative matrix factorization (NMF) and a standard BLSTM-RNN. Our proposed method outperforms the conventional methods and achieves an F1-score 74.9 % (error rate of 44.7 %) on the event-based evaluation, and an F1-score of 80.5 % (error rate of 33.8 %) on the segment-based evaluation, most of which also outperforms the best reported result in the DCASE 2016 task 2 challenge.

Index Terms— Polyphonic sound event detection, BLSTM-HMM, Sound activity detection, Hybrid system

1. INTRODUCTION

The goal of sound event detection (SED) is to identify the beginning and end of sound events and to identify and label these sounds. SED has great potential for use in many applications such as life-log, monitoring, environmental understanding, and automatic control of devices in a smart home. Improvements in machine learning techniques have opened new opportunities for progress in this challenging task. SED has thus been attracting more and more attention, and research in the field is becoming more active. Notably, some challenges such as the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [1, 2] have recently been held.

SED can be divided into two scenarios, monophonic and polyphonic. In monophonic SED, the maximum number of simultaneous active events is assumed to be one. On the other hand, in polyphonic SED, there can be any number of simultaneous active events. Polyphonic SED is a more realistic and difficult task than monophonic SED because in real-world situations, it is likely that several sound events will happen simultaneously, resulting in multiple sound events overlapping.

The most typical approach to SED is to use a hidden Markov model (HMM), where an emission probability distribution is represented by Gaussian mixture models (GMM-HMM), with mel frequency cepstral coefficients (MFCCs) as features [3, 4]. In the GMM-HMM approach, each sound event as well as silent regions is modeled by an HMM, and the maximum likelihood path is determined using the Viterbi algorithm. However, this approach shows limited performance, and requires heuristics such as the number of simultaneous active events to perform polyphonic SED. Another approach is to utilize non-negative matrix factorization (NMF) [5–8]. In the NMF approaches, a dictionary of basis vectors is learned by decomposing the spectrum of each single sound event into the product of a basis matrix and an activation matrix, then combining the basis matrices of all sound events. The activation matrix at test time is estimated using the combined basis vector dictionary, and used either for estimating sound event activations, or as a feature vector further passed to a classifier. These NMF-based methods show good performance, however, they do not take advantage of time axis information, and it is necessary to find the optimal number of bases for each sound events.

More recently, methods based on neural networks have also achieved good performance for SED [9–15]. In these neural network approaches, a single network was typically trained to be able to deal with a multi-label classification problem for polyphonic sound event detection. Furthermore, some studies [10, 12, 13, 15] have utilized recurrent neural networks (RNN), which are able to take into account correlations in the time direction. Although these approaches provide good performance, they perform frame-by-frame detection and do not have an explicit duration model for the output label sequence, and a threshold value for the actual outputs needs to be carefully decided in order to achieve the best performance.

In this study, we propose a hybrid system of HMM and bidirectional long short-term memory RNN (BLSTM-HMM[1]) system, where the output duration is controlled by an HMM on top of a BLSTM network, and extend the use of the hybrid system to polyphonic SED and, more generally, to the multi-label classification problem. Our approach not only allows to take advantage of time axis information and to introduce an explicit duration control, but also alleviates the need for thresholding and data dependent

[1]We submitted an initial version of this BLSTM-HMM system to the IEEE AASP Challenge DCASE 2016 [16]. We here further investigate the effect of the HMM transition probability, and propose an improved post-processing based on the SAD network, which achieved significant improvement from [16].

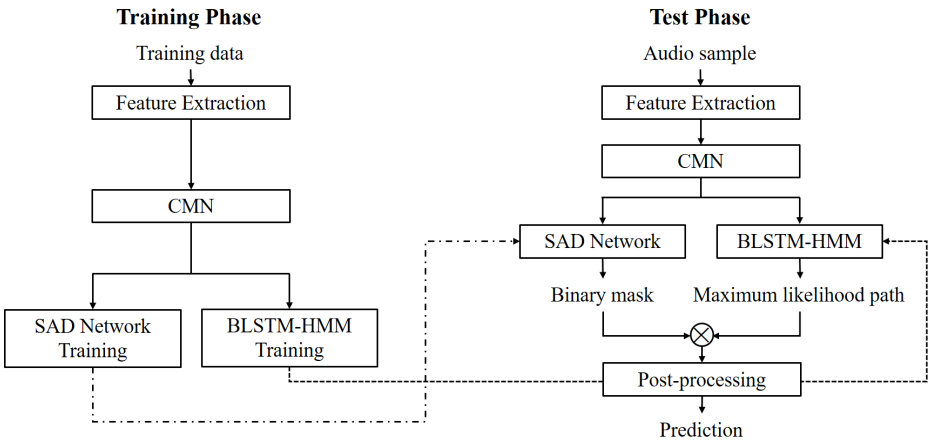


Fig. 1: System overview

processing. Furthermore, to effectively reduce insertion errors of low-volume sound events, which are often observed under noisy conditions, we additionally propose to perform binary mask post-processing using a sound activity detection (SAD) network which determines whether a segment is silent or contains an active sound event of any type, inspired by the well-known benefits of voice activity detection in speech recognition.

2. PROPOSED METHOD

2.1. System Overview

An overview of our proposed system, separated into training and test phases, is shown in Fig. 1, In the training phase, we extract the feature vectors and perform Cepstral Mean Normalization (CMN) for each training sample (Section 2.2). Using the obtained feature vectors, we train a BLSTM-HMM hybrid model (Section 2.3), and a sound activity detection (SAD) network (Section 2.4).

In the test phase, we extract the feature vectors from an input audio sample, and then perform CMN. The feature vectors are used as input into both the BLSTM-HMM and SAD networks. The BLSTM-HMM determines whether each sound event is active or not, while the SAD network estimates a binary mask which indicates global sound event activity, i.e., whether one or more sound events, whatever their types, are active in a given segment. Finally, we apply this binary mask to the activations of each sound event as estimated by the BLSTM-HMM (Section 2.4), and perform some more post-processing (Section 2.5).

2.2. Feature Extraction

The input signal is divided into 25 ms windows with 40 % overlap, and we compute 100 log-mel filterbank features for each window (we use more bands than usual since high frequency components are more important than low frequency ones for SED). After that, we perform cepstral mean normalization (CMN) for each piece of training data, thus obtaining the input feature vector $\mathbf{x}_t$ at frame t. These operations are performed using HTK [17].

2.3. BLSTM-HMM

We utilize the BLSTM-HMM hybrid model to capture sound event-dependent temporal structures and also to perform sequence-by-sequence detection without the thresholding as used in conventional frame-by-frame methods. We extend the hybrid model, which handles a multi-class classification problem in general, in order to handle a multi-label classification problem for polyphonic SED. In order to do this, we build a three state left-to-right HMM with a fourth non-active state for each sound event. The structure of our HMM is shown in Fig. 2, where $n = 0$, $n = 5$ and $n = 4$ represent the *initial state*, *final state*, and *non-active state*, respectively. Notice that the non-active state only pertains to the absence of activity of that particular event, and does not indicate whether other events are active or not. In this study, the transition probabilities are learned from the sequences of training data using the Viterbi training algorithm.

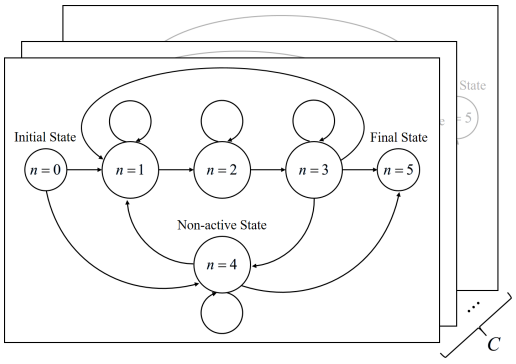
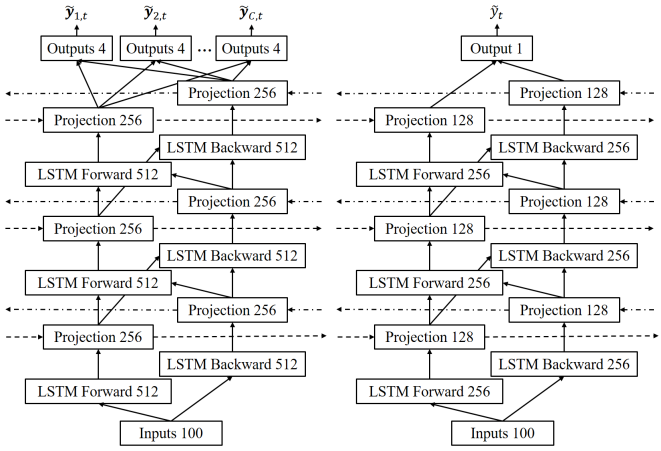


Fig. 2: HMM structure

In the BLSTM-HMM hybrid model, the BLSTM-RNN is used to calculate the HMM state posterior $P(s_{c,t} = n|\mathbf{x}_t)$, where $c \in \{1, 2, \dots C\}$ is the sound event index, $n \in \{1, 2, \dots, N\}$ the HMM state index, and $s_{c,t}$ the HMM state of event c at time t. From the HMM state posterior, the HMM state emission probability $P(\mathbf{x}_t|s_{c,t} = n)$ can be obtained using Bayes' theorem as follows:

$$P(\mathbf{x}_t|s_{c,t} = n) = \frac{P(s_{c,t} = n|\mathbf{x}_t)P(\mathbf{x}_t)}{P(s_{c,t} = n)}, \quad (1)$$

where the factor $P(\mathbf{x}_t)$ is irrelevant in the Viterbi computations. The structure of the network is shown in Fig. 3 (a). This network has three hidden layers which consist of a BLSTM layer with 1,024



(a) BLSTM-HMM network (b) SAD network

Fig. 3: Network structures

nodes, a projection layer with 512 nodes, and $C \times N$ output layer nodes. A softmax operation is used to ensure that the values of the posterior $P(s_{c,t}|\mathbf{x}_t)$ sum to one for each sound event c in frame t, as follows:

$$P(s_{c,t} = n|\mathbf{x}_t) = \frac{\exp(a_{c,n,t})}{\sum_{n'=1}^{N} \exp(a_{c,n',t})}, \tag{2}$$

where a represents the activation of the output layer node. The network is optimized by back-propagation through time (BPTT) with Stochastic Gradient Descent (SGD) and dropout using cross-entropy as shown by the following *multi-class, multi-label* objective function:

$$E(\mathbf{\Theta}) = \sum_{c=1}^{C} \sum_{n=1}^{N} \sum_{t=1}^{T} y_{c,n,t} \ln(P(s_{c,t} = n|\mathbf{x}_t)), \tag{3}$$

where $\mathbf{\Theta}$ represents the set of network parameters, and $y_{c,n,t}$ is the HMM state label obtained from the maximum likelihood path at frame t. (Note that this is not the same as the multi-class objective function in conventional DNN-HMM.) The HMM state prior $P(s_{c,t})$ is calculated by counting the number of occurrences of each HMM state. However, in this study, since our synthetic training data does not represent the actual sound event occurrences, the prior obtained from occurrences of HMM states has to be made less sensitive. Therefore, we smooth $P(s_{c,t})$ as follows:

$$\hat{P}(s_{c,t}) = P(s_{c,t})^{\alpha}, \tag{4}$$

where α is a smoothing coefficient. In this study, we set α to 0.01. Finally, we calculate the HMM state emission probability using Eq. 1 and obtain the maximum likelihood path using the Viterbi algorithm.

2.4. SAD Network

A common problem when performing polyphonic SED under noisy conditions is the decrease in performance due to insertion errors, where the background noise is misinterpreted as low-volume sound events. To solve this problem, we propose to perform binary masking using a sound activity detection (SAD) network. The SAD network identifies segments in which there is some sound event activity, whatever their type, similarly to voice activity detection (VAD) in the field of speech recognition. In this study, we train the network in Fig. 3 (b) as SAD network. This network has three hidden layers which consist of a BLSTM layer with 512 nodes, a projection layer with 256 nodes, and a single output layer node. The SAD network is optimized using BPTT with SGD and dropout under the following sigmoid cross-entropy objective:

$$E(\mathbf{\Theta}) = \sum_{t=1}^{T} y_t \ln(\tilde{y}_t) + (1 - y_t) \ln(1 - \tilde{y}_t), \tag{5}$$

where y_t is the reference data indicating presence or absence of sound events and $\tilde{y}_t$ the SAD network output.

We use a threshold of 0.5 to convert the SAD network outputs into a binary mask $\mathbf{M}$, and apply it to the activations $\mathbf{A}_c$ of each sound event c predicted by the BLSTM-HMM, as follows:

$$\tilde{\mathbf{A}}_c = \mathbf{M} \odot \mathbf{A}_c. \tag{6}$$

Note that the same binary mask $\mathbf{M}$ is applied to the activation of each sound event, and that the binary mask only has an effect on the insertion of sound events, not on the substitution or deletion of sound events.

Table 1: Experimental conditions

Sampling rate	44,100 Hz
Window size	25 ms
Shift size	10 ms
# training data	4 s $\times$ 100k samples
# development data	120 s $\times$ 18 samples
# evaluation data	120 s $\times$ 54 samples
# sound event classes	11
Learning rate	0.0005
Initial scale	0.001
Gradient clipping norm	5
Batch size	64
Time steps	400
# epochs	20

2.5. Post-processing

After masking, we perform three kinds of post-processing: 1) applying a median filter with a predetermined filter span, 2) filling up gaps which are shorter than a predetermined number, 3) removing events which are too short in duration. We set the median filter span to 170 ms (= 17 frames), the acceptable gap length to 1 s (= 100 frames), and the minimum duration threshold of each sound event to half of the minimum duration for that sound event as calculated from training data.

3. EXPERIMENTS

3.1. Experimental Conditions

We evaluate our proposed method by using the DCASE 2016 task 2 dataset [2]. The dataset includes a training set consisting of 20 clean audio files per event class, a development set consisting of 18 synthesized audio files of 120 sec in lengths, and an evaluation set consisting of 54 synthesized audio files of the same length as the development set files. The number of sound event classes in the dataset

Table 2: Experimental results

	Event-based (dev / eval)		Segment-based (dev / eval)	
	F1 [%]	ER [%]	F1 [%]	ER [%]
NMF (Baseline)	31.0 / 24.2	148.0 / 168.5	43.7 / 37.0	77.2 / 89.3
BLSTM-RNN	69.9 / 60.1	73.2 / 91.2	87.2 / 77.1	25.8 / 44.4
+ post-processing	81.5 / 71.2	35.7 / 50.9	89.3 / 79.0	20.3 / 36.8
+ SAD binary masking	82.2 / 73.7	34.0 / 45.6	89.5 / 79.9	19.8 / 34.3
BLSTM-HMM	77.4 / 67.9	46.5 / 64.3	87.7 / 78.8	23.3 / 40.3
+ trans. learning	80.0 / 71.0	38.6 / 55.1	88.8 / 79.6	20.4 / 37.4
+ post-processing	81.3 / 71.7	35.2 / 52.3	89.1 / 79.5	19.4 / 36.7
+ SAD binary masking	**82.6 / 74.9**	**32.7 / 44.7**	**89.7 / 80.5**	**18.3 / 33.8**

is 11. The development set is synthesized using the training set, and the evaluation set is synthesized using unknown samples.

For this study, we chose to further split the training data to build an open condition development set, which is lacking in the original dataset: we randomly selected 5 samples per event from the training set, and generated 18 samples which have 120 s length to be similar to the DCASE 2016 task 2 development set. These generated samples are used as development data to check the performance in open conditions. We used the remaining 15 samples per class to build our own training data. Instead of simply using the corresponding original training data, which is too small for training an RNN with sufficient depth, we performed training data augmentation by synthetically generating our own training data using the clean sound event samples and background noise as follows: 1) generate a silence signal of a predetermined length, 2) randomly select a sound event sample, 3) add the selected sound event to the generated silence signal at a randomly selected location, 4) repeat Steps 2 and 3 a predetermined number of time, 5) add a background noise signal at a predetermined signal to noise ratio (SNR). We set the signal length to 4 seconds, the number of events to a value from 3 to 5, the number of overlaps to a value from 1 to 5, SNR to a value from -9 dB to 9 dB, and finally synthesize 100,000 samples (= 111 hours).

Evaluation is conducted in two regimes, *event-based* (onset-only) and *segment-based* evaluation, where the F1-score (F1) and the error rate (ER) are utilized as evaluation criteria (see [18] for more details). All networks are trained using the open source toolkit TensorFlow [19] with a single GPU (Nvidia Titan X). Details of the experimental conditions are shown in Table 1.

3.2. Experimental Result

To confirm the performance of our proposed method, we compare it with two conventional methods, NMF (DCASE 2016 task 2 baseline) and standard BLSTM-RNN. NMF is trained on 15 clean samples per class using the DCASE 2016 task 2 baseline script [2]. BLSTM-RNN has the same network structure as in Fig. 3 (a) with the exception that the output layer is replaced by C nodes with sigmoid activation function, one node for each of the C sound events. Each node's output $y_c \in [0, 1]$ is binarized to determine event activity. We set the threshold to 0.5, i.e., sound event c is considered active if $y_c > 0.5$, and inactive otherwise.

Experimental results are shown in Table 2. Our proposed system outperformed conventional methods on both of evaluation criteria, and this is the best performance in the DCASE 2016 task 2 challenge [2] except for segment-based error rate. From these results, we can see that the it is important for polyphonic SED to capture the sound event-dependent temporal structures.

Next, we focus on the effect of transition probability and post-processing. In the study [16], we used a fixed transition probability because we expected the emission probability calculated by the BLSTM to be dominant regarding the decision of the maximum likelihood path. However, the results of the current study show that even if we use a hybrid model of neural network and HMM, performance can be improved by using an appropriate transition probability. Our results also show the effectiveness of the proposed post-processing for both BLSTM-RNN and BLSTM-HMM, which we did not perform in [16]. The most effective form of post-processing is the removal of events which have too short duration for their particular type, because each sound event typically has a characteristic duration which is different from that of other events. These observations point to the importance of explicit duration control (e.g., hidden semi-Markov model (HSMM) [20, 21]).

Finally, our results confirm the effectiveness of our proposed SAD masking for both BLSTM-RNN and BLSTM-HMM, especially on the reduction of the error rate for the evaluation set. This is because the background noise in the evaluation set is noisier than in the development set, therefore leading to many insertion errors.

4. CONCLUSION

In this study, we proposed a new method for polyphonic SED based on a hybrid bidirectional long short-term memory/hidden Markov model system (BLSTM-HMM) combined with a sound activity detection (SAD) network. The BLSTM-HMM not only provides an explicit duration model for output labels, but also alleviates the need for thresholding, which outperformed conventional methods on both of evaluation criteria. Furthermore, binary masking using an SAD network prevents the decrease in performance caused by insertion errors under noisy conditions, which improved the performance of all experimental configurations.

In future work, we will investigate the reason for decreased performance in the segment-based evaluation for BLSTM-HMM, the use of a more flexible duration control model such as HSMM, and the application of our proposed method to a real-recording dataset.

5. REFERENCES

[1] "Detection and Classification of Acoustic Scenes and Events 2013 - DCASE 2013," http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/.

[2] "Detection and Classification of Acoustic Scenes and Events 2016 - DCASE 2016," http://www.cs.tut.fi/sgn/arg/dcase2016/.

[3] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.(WASPAA)*, 2013.

[4] 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, pp. 1–13, 2013.

[5] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen, "Sound event detection in multisource environments using source separation," in *Workshop on machine listening in Multisource Environments*, 2011, pp. 36–40.

[6] S. Innami and H. Kasai, "NMF-based environmental sound source separation using time-variant gain features," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1333–1342, 2012.

[7] A. Dessein, A. Cont, and G. Lemaitre, "Real-time detection of overlapping sound events with non-negative matrix factorization," in *Matrix Information Geometry*, pp. 341–371. Springer, 2013.

[8] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 45–49.

[9] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

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

[11] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1, 2015.

[12] Y. Wang, L. Neves, and F. Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2742–2746.

[13] F. Eyben, S. Böck, B. Schuller, A. Graves, et al., "Universal onset detection with bidirectional long short-term memory neural networks.," in *ISMIR*, 2010, pp. 589–594.

[14] I. Choi, K. Kwon, S. H. Bae, and N. S. Kim, "DNN-based sound event detection with exemplar-based approach for noise reduction," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 16–19.

[15] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 6–10.

[16] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 35–39.

[17] "HTK Speech Recognition Toolkit," http://htk.eng.cam.ac.uk.

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

[19] "TensorFlow - An open source software library for machine intelligence," https://www.tensorflow.org.

[20] S. Z. Yu, "Hidden semi-markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010.

[21] Z. Heiga, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A hidden semi-markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 825–834, 2007.