Early detection of continuous and partial audio events using CNN

Ian McLoughlin[1,2], *Yan Song*[2], *Lam Pham*[1], *Ramaswamy Palaniappan*[1], *Huy Phan*[3], *Yue Lang*[4]

[1]The University of Kent, School of Computing, Medway, UK
[2]The University of Science and Technology of China, Hefei, PRC
[3]University of Oxford, Department of Engineering Science, Oxford, UK
[4]Huawei European Research Center, Munich, Germany

ivm@kent.ac.uk, songy@ustc.edu.cn, ldp7@kent.ac.uk, pr254@kent.ac.uk,
huy.phan@eng.ox.ac.uk, langyue@huawei.com

Abstract

Sound event detection is an extension of the static auditory classification task into continuous environments, where performance depends jointly upon the detection of overlapping events and their correct classification. Several approaches have been published to date which either develop novel classifiers or employ well-trained static classifiers with a detection front-end. This paper takes the latter approach, by combining a proven CNN classifier acting on spectrogram image features, with time-frequency shaped energy detection that identifies seed regions within the spectrogram that are characteristic of auditory energy events. Furthermore, the shape detector is optimised to allow early detection of events as they are developing. Since some sound events naturally have longer durations than others, waiting until completion of entire events before classification may not be practical in a deployed system. The early detection capability of the system is thus evaluated for the classification of partial events. Performance for continuous event detection is shown to be good, with accuracy being maintained well when detecting partial events.

Index Terms: sound event detection, convolutional neural networks, audio classification, segmentation.

1. Introduction

Continuous sound event detection means the identification of sound events as they occur in a continuous audio medium. It extends the classification of isolated and separated sounds into real-world machine hearing scenarios. This is important for smart home and vehicle environments, speech interaction and telecommunication systems, and has relevance to audio-based security monitoring, ambient event detection and auditory scene analysis. Sound event detection research has traditionally been driven by techniques developed for speech recognition, including Mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLPs) with Gaussian mixture models (GMMs) and hidden Markov models (HMMs) [1, 2, 3, 4, 5]. However these features and methods have more recently been surpassed by spectrogram-based techniques [6, 7], especially for the classification of noise-corrupted sounds. Recent systems have demonstrated very good results from the use of deep learning, including deep neural networks (DNN) [8, 9, 10, 11] and convolutional neural networks (CNN) [12, 13]. Both DNN and CNN classifiers perform well in the presence of acoustic background noise, with the latter demonstrating superior noise robustness.

While acoustic noise robustness is an important real-world attribute of such systems, practical methods must also have the capability to distinguish between the absence of sound events, the presence of individual events, and the occurrence of overlapping events, and do so in levels of signal-to-noise (SNR) that are unknown a priori. The task is particularly difficult when many possible sound classes are involved, and when some classes have an inherently noise-like sound.

This paper proposes a detection front-end to identify seed regions from spectrogram image features that have the characteristic time-frequency shape of sound events, prior to classification. Detected seed regions are then classified using a well-trained CNN to classify zero, one or multiple events. The seed region detector is further optimised to enable early event detection. This is inspired by systems such as [14, 15] which aim to enable reliable classification of sound events as they are occurring, rather than waiting until they have completed (i.e. online classification). This is an important requirement for future real-time machine hearing systems that need to classify sound events that have long durations. We evaluate performance on the standard continuous audio event detection task first developed in [16] and extended in [17], then evaluate the abilities of the system when forced to perform partial detection. Results show very good performance for full event detection, and gracefully degrade as classification is performed earlier.

2. Background

The basic classifier in many recent sound event classifiers is typically trained in a supervised fashion using data which is presented in individual files. Each file contains an isolated sound event without added noise, corresponding to a single class. In the baseline CNN classifier used in this paper (in Section 3), spectrogram image features (SIF) are obtained from individual labelled sounds, conditioned, downsampled, and used to train a CNN. Since the training material has no added background noise, a basic energy detector is easily capable of identifying regions of interest in the SIF prior to training.

For classification of detected sounds, many types of feature have been explored in the research literature, including raw waveform, MFCC, several kinds of spectrogram and correlogram, as have many kinds of back-end classifier. For example MFCC-HMM [18], SIF-SVM [8], SIF-DNN [8] and SIF-CNN [12]. Each of those systems was evaluated in clean and noisy isolated sounds (known as robust sound event classification), using a standard 50-class evaluation of real-world sounds first proposed by Dennis [18]. However real-world audio is continuous rather than discrete, with sounds of unknown duration occurring at unknown, perhaps overlapping, times. A *detection* operation is thus required in conjunction with the *classification* task.

For this reason, an experimental evaluation was proposed by the authors [17] that combined detection and classification of real-world sounds in continuous waveforms that included overlapping sounds – with the test material as illustrated in Fig. 1. The system proposed and evaluated in this paper for the robust classification of continuous and overlapping sounds, uses identical training data, but enhances the evaluation further through the development of early-detection capabilities, inspired by those first introduced in [19].

Early detection is another capability that is important for real-world sound event detection. Some sound events have longer durations than others, and waiting until completion of entire sound events before classification – as most current systems do (including [17]) may be impractical for longer sounds. Early detection is needed for *online detection*, and the degree of earliness is a factor in the classification latency of a system.

3. The proposed detection system

The proposed system is shown in Fig. 2, roughly divisible into the detection process (top half) and the classifier (bottom half). Within the classifier, a CNN architecture is employed that is unchanged from the baseline classifier in [17]; this means that improvements in performance are due to the capabilities of the detection system alone.

3.1. Spectrogram image features

Both DNN and CNN classifiers have been shown very capable of extracting discriminative information from spectrogram features [8, 12, 13], with the best performing classifiers being CNN-based, and acting on SIF features. the SIF extraction process is; (a) take FFT magnitude of overlapping analysis windows (size 25 ms, overlap 20 ms), (b) downsample in both time and frequency to a 52×40 patch, (c) normalise in amplitude and (d) optionally denoise prior to classification.

3.2. Energy detector

During training – which uses clean and labelled sound files – energy gating is used to select SIF patches for classification, with up to 9 patches per file (with one sound per file and 2500 files in total) being used to contribute to the training. While this works well when testing clean sounds, the method is easily defeated by background noise, and it does not work well for overlapping sounds or complex multi-part sounds. More noise-robust methods are thus required for testing.

Waveform frames are processed sequentially from each sound file during training, with up to 9 highest energy frames and their immediate 40-frame context being selected as an image patch. A hold-off of 20 frames is imposed until the next patch can be selected, and frames with energy lower than 10% of peak energy inside the context are excluded. This confers a degree of noise resistance, with the hold-off period designed to ensure that loud sounds spanning multiple frames do not dominate over quieter sounds occurring elsewhere. This applies to sounds characterised by a strong attack energy and a sustained release, or multi-part sounds that have double or multiple energy peaks (e.g. *stapler*, *footsteps*, *doorbell*).

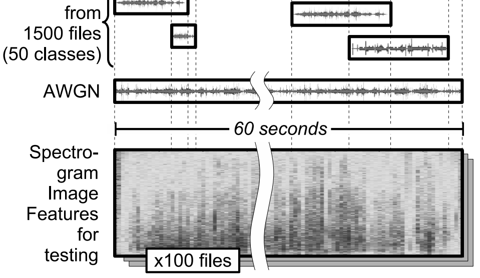
15 sounds from 1500 files (50 classes)
AWGN
60 seconds
Spectrogram Image Features for testing
x100 files

Figure 1: *Illustration of the continuous test material.*

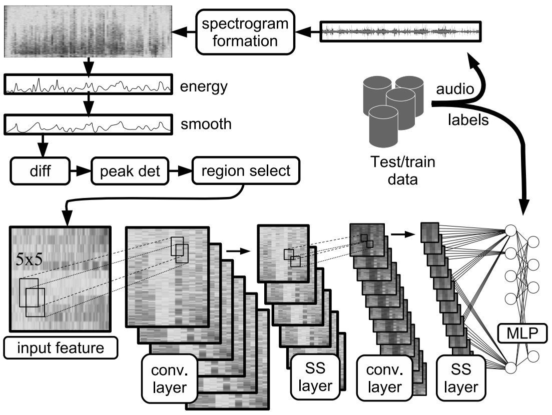

Figure 2: *Block diagram of the classifier in test mode.*

The CNN classifier outputs posterior probability P_k, for each image patch, over $k = 1...50$ classes. Index $n = \arg\max(P_k), k = 1...50$ identifies the highest probability class, but is only accepted if $P_n > P_{th}$, otherwise this sound event is classed as noise. As mentioned above, this energy-gated detector is used primarily during training.

3.3. Shape-based seed detection

Discrete sound events in nature are characterised by their acoustic energy, which is often the result of the conversion of kinetic energy to sound, where the cause is percussive or frictional, or the resonance of moving air (which itself is the conversion of kinetic energy in the air to correlated wave motion). The observation of the authors is that the physical basis for sound creation means that sound energy from single events tends to be either narrowband in frequency yet of relatively long duration, or is wideband but of shorter duration. Percussive sounds, clicks, staplers, claps and bangs have wideband, short duration energy releases. Horns, whistles, bells, squeaks typically have narrowband acoustic releases, but of longer duration. Even if the same amount of energy is generated/received for each sound, its shape in the time-frequency space will differ. This observation motivated the creation of a shape-based detector that detects either narrow-but-long or wide-but-short regions.

In operation, the detector computes the energy from the spectrogram, $\mathcal{S}$ of L_x frequency bins over L_y frames, i, so that; $E_i = \sum_{y=0}^{L_y} |\mathcal{S}(i,y)|$ and then the box filter-smoothed envelope $\widetilde{E}_i = \sum_{l=1}^{P} a_l.E_{i-l}$ is extracted, where $a_l = 1$ for $0 < l < 240$. Peak candidates are obtained from the differential of the envelope, $\widetilde{E}'$ and then sorted by peak energy with a 240-frame hold-off and a minimum height threshold of 1.0. Energy is computed over a longer time span, encouraging both short duration, wideband energy events, as well as longer narrowband events (i.e. instantaneous frame energy is unimportant). Thresholding then improves noise rejection, similar to the thresholding mentioned in Section 3.2, P_{th} defines the

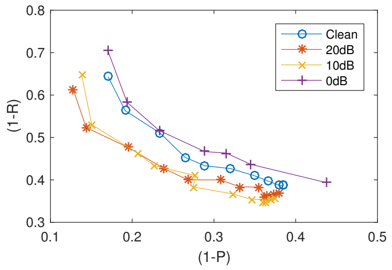

Figure 3: *Det curve for energy-only front-end detector operating on clean, 20dB, 10dB and 0dB SNR sound recordings.*

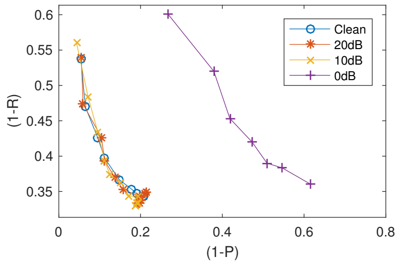

Figure 4: *Det curve for shape-based front-end detector operating on clean, 20dB, 10dB and 0dB SNR sound recordings.*

minimum probability threshold for detection of any class at the output of the classifier (we sweep P_{th} during experiments, but the best results are generally obtained when $P_{th} \approx 0.05$).

3.4. Convolutional neural network

CNNs are well known in image classification [20, 21], and in this application, the spectrogram patch is an image. The CNN structure used, derived from [17], can be seen in Fig. 2. It has 5-layers (2 convolutional layers, 2 subsampling layers and 1 fully connected layer), with $52 \times 40 = 2080$ input dimensionality and 50 output classes from a single fully connected layer. The first and second convolutional layers consist of 6 and 12 kernels, each with a kernel size of 5×5. The subsampling layers employ average pooling with a common factor of 2:1. Batch normalization [22] is applied before each convolutional layer.

3.5. The evaluation task

The sound material used for training and evaluation consists of 4000 recordings divided into 50 different sound event classes, each of 80 files. The files were randomly selected from the Real World Computing Partnership (RWCP) Sound Scene Database in Real Acoustic Environments [23] across a subset of 50 classes, as specified in [7]. Of the 80 files in each class, 50 were randomly selected to be the training set $(50 \times 50 = 2500)$ with the remainder $(30 \times 50 = 1500)$ being used for evaluation.

The evaluation material is formed by first creating 100 separate 1-min long empty test files to which 15 randomly-selected test sound events are inserted at random time indices. The random nature of the selection means that some sounds are represented multiple times per test file, and that double and even triple overlap events occur. In the original definition of the evaluation method [17], noise was randomly selected from random positions within four different NOISEX-92 noises, however the tests in the current paper employ only AWGN, which improves the repeatability of the experiments. One further change is made to the current evaluation compared to the testing methodology described in [17]. The is the adoption of a much stricter criterion for class detection; in the current paper, *any* analysis frame containing *any* classes with posterior probability exceeding P_{th} are counted as a detection, with the detection being correct only if the candidate classes match the ground truth. There may be between 0 and many (up to 50 if P_{th} is low) detections per analysis frame, and perhaps several hundred analysis frames for each ground truth class region. Yet each ground truth class region can only contribute either 0 or 1 correct detections. In the previous work [17], detections were made for each analysis frame in the same way, but correct detections were counted for each analysis segment, rather than for a whole ground truth class region. Therefore, it was possible that there could be many correct detections within a single ground truth class region (e.g. if one ground truth class region contained 20 analysis segments, there could be up to 20 correct detections counted across that region, rather than up to 1 in the current system). The stricter criterion is important because we are measuring early detection, which affects event-based detection much more than frame-based detection. We therefore first re-evaluate the baseline detector from [17] using the stricter criterion.

In the reported results, we define precision as $P = M/N$, where M is the number of ground truth sound events detected correctly, and N is the total number of detected events. Recall is computed as $R = M/K$, where K is the total number of ground truth sound events in the test. The composite F_1 score combines both metrics to yield a single overall performance figure; $F_1 = 2/(P^{-1} + R^{-1})$.

4. Results and discussion

4.1. Energy and shape detection

We first explore the performance of the system with a basic energy detector. Fig. 3 plots the recall against precision for a range of P_{th} thresholds in clean and noisy conditions. The results show degradation in overall detection and classification performance due to the presence of noise. This is not unexpected, given that region detection is based only on patch energy. The shape-based detector of Section 3.3 was then applied and the above tests repeated, with results plotted in Fig. 4. In this case, very little degradation was experienced at 20dB SNR, or even at 10dB SNR, although at 0dB SNR it is significantly degraded.

Further results are given in Table 1. Results were obtained for a range of peak candidate thresholds P_{th} around the maximal F_1 region, and the scores at which peak F_1 occurs are reported for each test. For now, consider just the lines beginning with “full”, which are the results in which early detection is not being evaluated.

It is interesting to note that the highest F_1 score actually occurs when low levels of noise are present – due to the fact that even ‘clean’ recordings contain low levels of noise, and that it is better to spread noise evenly than to cluster it around sound events. The same phenomenon was found in CNN classification of isolated sounds (e.g. in [12]) where low levels of background noise tended to be beneficial to performance. Nevertheless, as noise increases beyond 10dB SNR, performance degrades, so that scores at 0dB are very poor, in common with prior methods such as [17]. Even with isolated sound event classification [8], recognition of sounds in 0dB SNR is extremely challenging. From the overall results presented so far, the best F_1 for each

Table 1: *Precision (P), recall (R) and F_1 score for the original energy-based detector and the proposed shape-based detector performing feature selection with backend CNN-based classification. The results report the best achieved F_1 score over a P_{th} range [0.01:0.95] with a step size of 0.05 for clean, 20dB, 10dB and 0dB SNR AWGN and early detection degrees of 100%, 50%, 25% and 12.5%.*

	Clean			**20dB**			**10dB**			**0dB**		
Earliness	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Energy-based detector												
full	0.711	0.567	0.631	0.732	0.600	0.659	0.725	0.617	0.667	0.711	0.533	0.610
50%	0.711	0.517	0.598	0.749	0.587	0.658	0.763	0.580	0.659	0.798	0.487	0.605
25%	0.667	0.373	0.479	0.776	0.403	0.531	0.740	0.473	0.577	0.511	0.403	0.451
12.5%	0.084	0.057	0.068	0.135	0.073	0.095	0.127	0.083	0.101	0.025	0.077	0.038
Shape-based detector												
full	0.852	0.633	0.727	0.843	0.647	0.732	0.814	0.670	0.735	0.582	0.547	0.564
50%	0.839	0.627	0.718	0.851	0.630	0.724	0.870	0.623	0.726	0.633	0.540	0.583
25%	0.750	0.490	0.593	0.790	0.477	0.595	0.786	0.490	0.604	0.659	0.450	0.535
12.5%	0.376	0.137	0.200	0.373	0.137	0.200	0.361	0.143	0.205	0.292	0.150	0.198

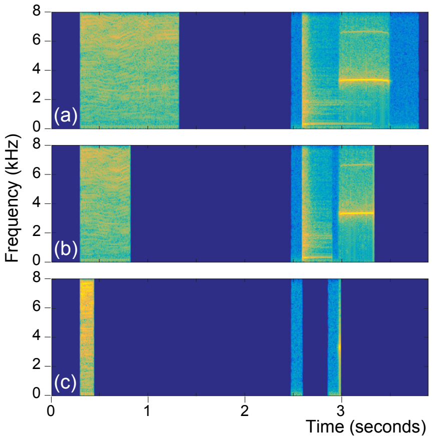

Figure 5: *Spectrogram of a fixed of one of the 100 test files from the (a) full, (b) 50% and (c) 12.5% event test databases.*

noise condition for the shape-based detector system compares well with the energy-based detector apart from in 0dB AWGN.

4.2. Early detection

Early detection was then explored by creating four sets of experimental continuous sound recordings. Each used the same random selection of sound events, starting positions and overlaps, but only included the beginning segment of each sound included in the test. It is thus a task of detecting partial sounds, but since these segments all include the beginning of the sounds under question, with the end truncated, it forces the system to perform detection on just the early parts of each sound. The task is illustrated in Fig. 5, which shows a fixed short segment of spectrogram from a single experimental condition, from three early detection databases. In each case, these are clean sounds without additional AWGN. In the figure, the same three events are present, starting at the same position in each recording. The full event data (a) includes the entire sound for each of the three events, whereas in (b) only the first half of each sound has been included, and in (c) only the first 12.5% has been pasted in. The 25% data has not been shown for space reasons, but follows a similar pattern. For each of the experiments, classification uses this data alone with no a priori information regarding the length of each event. It is interesting to note that as the length of event is curtailed, the degree of overlap also reduces; the full data test in spectrogram (a) has a significant overlap between the second and third events. The overlap is small when only 50% of the sounds are included, and is absent in the 12.5% case (although overlap still occurs in other parts of the test database).

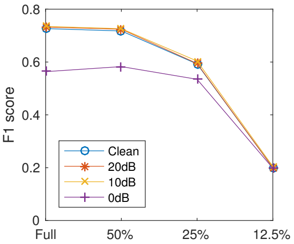

Figure 6: *F_1 score achieved by the shape-based detector in different levels of AWGN, for each early detection condition.*

Full results for precision, recall and F_1 score are presented in Table 1 for both energy-based and shape-based detector, for each early detection condition. The shape-based detector results degrade much less for the early detection cases than do the energy-based detector results. In fact, degradation due to early event detection is small up to even 25%, and may even be beneficial in some cases (for example, some slightly improved accuracy for 50% early detection), which we believe is due to a trade off between less data being available for classification, and the reduction in overlap. Fig. 6 shows the peak F_1 score for each tested condition of the shape-based detector.

5. Conclusion

This paper has proposed a shape-based front-end detector that operates in conjunction with a well-trained isolated sound CNN classifier, to perform robust early sound event detection. The baseline CNN classifier is first evaluated in clean and noisy conditions, using a standard acoustic noise database, with a simple energy-based front-end. The proposed shape-based detector is then evaluated in the same conditions, and shown to improve performance. The early-detection task is derived from the standard test methodology, allowing performance to be evaluated for four early-detection conditions. The new detector allied with the backend CNN classifier are shown to perform very well when even 50% of each sound event is omitted, and to degrade gracefully as detection is forced on the basis of less and less classification data.

6. References

[1] H. Phan, M. Maas, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 1, pp. 20–31, 2015.

[2] H. Phan, L. Hertel, M. Maass, R. Mazur, and A. Mertins, "Learning representations for nonspeech audio events through their similarities to speech patterns," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 807–822, April 2016.

[3] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro, "Non-speech audio event detection," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 1973–1976.

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[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] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 130–133, 2011.

[7] J. Dennis, H. D. Tran, and E. S. Chng, "Image feature representation of the subband power distribution for robust sound event classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 367–377, 2013.

[8] I. McLoughlin, H.-M. Zhang, Z.-P. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 540–552, Mar. 2015.

[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)*, July 2015, pp. 1–7.

[10] T. L. Nwe, T. H. Dat, and B. Ma, "Convolutional neural network with multi-task learning scheme for acoustic scene classification," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 1347–1350.

[11] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "A comparison of deep learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 126–130.

[12] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, no. 2635. IEEE, Apr. 2015, pp. 559–563.

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

[14] H. Phan, M. Maass, R. Mazur, and A. Mertins, "Acoustic event detection and localization with regression forests," in *15th Annual Conference of the International Speech Communication Association (Interspeech)*, Singapore, September 2014, pp. 1–5.

[15] H. Phan, P. Koch, I. McLoughlin, and A. Mertins, "Enabling early audio event detection with neural networks," *arXiv preprint arXiv:1712.02116*, 2017.

[16] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event detection in continuous audio environments," in *Proc. Interspeech*, Sep. 2016.

[17] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, and H. Phan, "Continuous robust sound event classification using time-frequency features and deep learning," *PloS one*, vol. 12, no. 9, p. e0182309, 2017.

[18] J. W. Dennis, "Sound event recognition in unstructured environments using spectrogram image processing," Ph.D. dissertation, Nanyang Technological University, Singapore, 2014.

[19] H. Phan, M. Maass, R. Mazur, and A. Mertins, "Early event detection in audio streams," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, June 2015, pp. 1–6.

[20] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML*, 2015, pp. 448–456.

[23] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, "Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition," in *EUROSPEECH*, 1999, pp. 2255–2258.