INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Deep Spectral–Cepstral Fusion for Shouted and Normal Speech Classification

Takahiro Fukumori

College of Information Science and Engineering, Ritsumeikan University, Japan
fukumori@fc.ritsumei.ac.jp

Abstract

Discrimination between shouted and normal speech is crucial in audio surveillance and monitoring. Although deep neural networks are used in recent methods, traditional low-level speech features are applied, such as mel-frequency cepstral coefficients and the mel spectrum. This paper presents a deep spectral–cepstral fusion approach that learns descriptive features for target classification from high-dimensional spectrograms and cepstrograms. We compare the following three types of architectures as base networks: convolutional neural networks (CNNs), gated recurrent unit (GRU) networks, and their combination (CNN–GRU). Using a corpus comprising real shouts and speech, we present a comprehensive comparison with conventional methods to verify the effectiveness of the proposed feature learning method. The results of experiments conducted in various noisy environments demonstrate that the CNN–GRU based on our spectral–cepstral features achieves better classification performance than single feature-based networks. This finding suggests the effectiveness of using high-dimensional sources for speech-type recognition in sound event detection.

Index Terms: shouted and normal speech classification, spectral-cepstral fusion, deep neural network

1. Introduction

Audio surveillance systems [1, 2], which detect abnormal situations using microphones, have garnered significant attention as systems that enhance safety in daily life. Classifying an audio segment into shouted and normal speech is crucial to facilitate emergency rescue. Recently, deep learning methods have been used in this classification task, e.g., convolutional neural networks (CNNs) and recurrent neural networks were used to model the relationship between the temporal variation of speech features and speech status [3–5]. Most conventional studies, including state-of-the-art studies [5, 6], used traditional speech features, such as mel-frequency cepstral coefficients (MFCCs).

Voiced speech is generally generated by the vibration of vocal folds that create periodic excitations to the vocal tract during the pronunciation of phonemes, and MFCCs are designed to represent such vocal tracts in the cepstral domain. However, other aspects should also be considered. For example, the vocal folds as well as ventricular folds of a person vibrate strongly when he/she is shouting [7], and the duration at the end of the word typically becomes longer than that of a normal speech [8]. MFCCs might not exhibit these cepstral characteristics owing to dimensionality reduction during feature extraction. Valenzise et al. [9] showed that the harmonic energy of shouted speech is concentrated in a frequency band ranging between 1 and 2.5 kHz, implying the importance of spectral features. Hence, we should learn features from both cepstral and spectral domains, as they are descriptive for target classification and do not rely on conventional hand-crafted speech features.

Herein, we present a novel deep neural network (DNN) for shouted and normal speech classification that complementarily uses features from spectral and cepstral domains. We compare the following three types of network structures based on using single features: the CNN, gated recurrent unit (GRU) network [10], and their combination, CNN–GRU. The proposed DNN comprises two single feature-based networks for combining the features of the two domains for classification. The experimental results show that our network, whose inputs are both spectrograms and cepstrograms, achieved better classification performance than the conventional low-level features or networks that use only a single domain.

The main contributions of this study are summarized as follows: (i) We propose a novel approach for classifying shouted and normal speech, where features are learned from raw, high-dimensional spectral and cepstral information; (ii) using a corpus comprising real shouts and normal speech, we provide a comprehensive performance evaluation by changing the combinations of features, DNN architectures, and noise conditions.

2. Related Studies

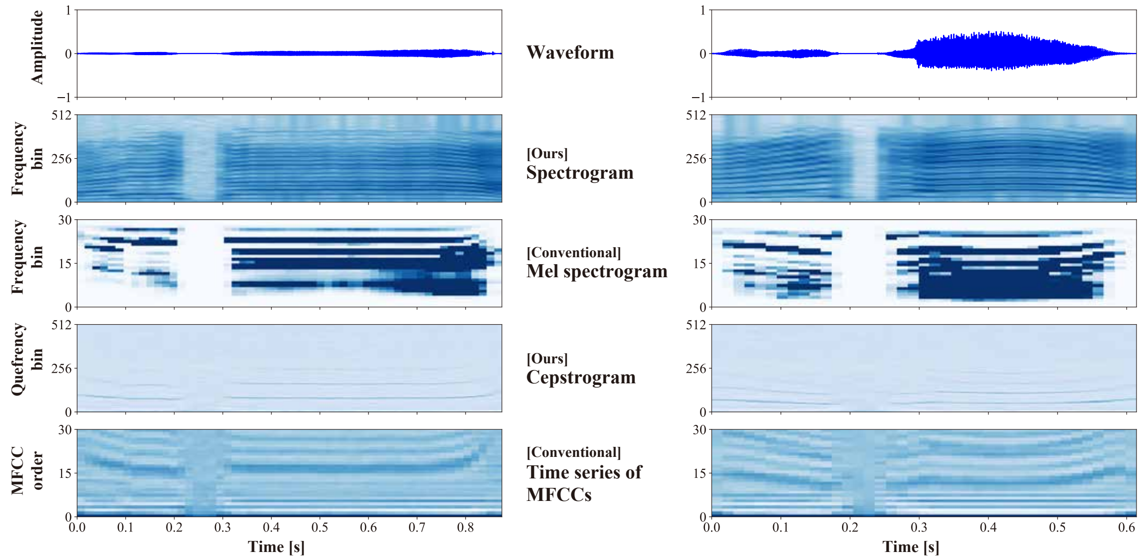
In earlier studies regarding shouted and normal speech classification, MFCCs were primarily used to train classifiers, such as Gaussian mixture models [11, 12], support vector machines [13, 14], and hidden Markov models [15, 16]. Other well-known features include spectral features such as pitch, harmonic-to-noise ratio, spectral centroid, flux, and flatness, all of which are scalar values. The recent progress in DNNs is evident in shouted speech detection. Laffitte et al. [17] presented a pioneering study based on a deep belief network model, which used MFCCs of successive frames as inputs. Gaviria et al. [4] recently used MFCCs and mel-spectrograms in a DNN framework to classify shouted speech. Although recent studies regarding speech enhancement [18] and emotion recognition [19] demonstrated the effectiveness of raw waveform-based representation learning, studies that use both spectrograms and cepstrograms to train DNNs have not been reported.

Our task can be regarded as a subtopic of sound event detection, which has been actively investigated in recent years [20–23]. There are various benchmark sets and competitions for sound event detection that provide predefined categories. However, none of these tasks contain both normal and shouted speech categories. In this study, we provide insights into learning effective features for such speech-type categorization.

3. Methodology

This section presents a deep spectral–cepstral fusion approach for shouted and normal speech classification. Section 3.1 describes the speech features in spectral and cepstral domains, which are used in both conventional methods and the proposed method, respectively. Section 3.2 provides the details of DNN architectures whose inputs are single features. Our method concatenates the outputs of single DNNs to yield a classification result, which is presented in Section 3.3.

http://dx.doi.org/10.21437/Interspeech.2021-1245



(a) *Normal speech.* (b) *Shouted speech.*

Figure 1: *Examples of waveforms of a female speaker's normal and shouted speech and their corresponding speech features.*

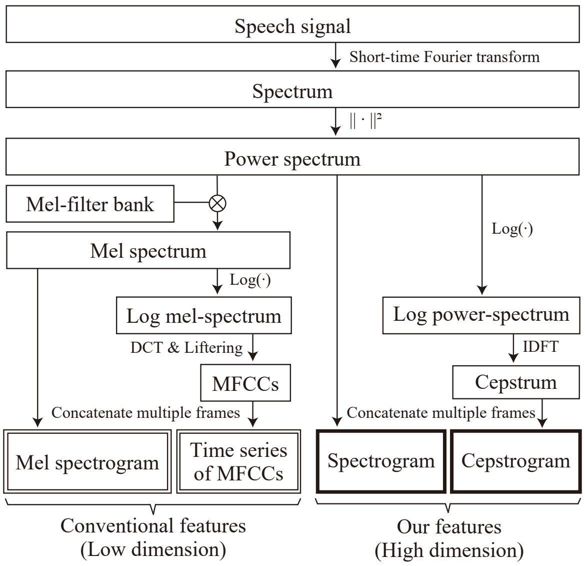


Figure 2: *Extraction of conventional features and our high-dimensional features.*

3.1. Speech feature extraction

For a specified audio segment, we partitioned it into successive frames using a Hamming window that had a length of 1,024 points (i.e., 64 ms) and a hop length of 512 points (i.e., 32 ms); subsequently, we obtained the features for every 20 frames. Figure 1 shows the spectrogram, mel-spectrogram, cepstrogram, and MFCCs, which were extracted from waveforms in which a female speaker uttered and shouted. Figure 2 summarizes the extraction of these speech features. First, we provide the details of the MFCCs and mel-spectrogram, which are used in conventional methods [4–6, 11–17, 24–27] (hereinafter, **conventional low-level features**).

Time series of MFCCs: MFCCs are typical cepstral features. Most conventional methods of shouted speech detection use MFCCs with dimensions ranging between 8 and 60 [5, 6, 11–17, 24–27]. Following [11], we extracted 30-dimensional MFCCs from each frame and concatenated vectors over 20 frames, resulting in a 600-dimensional cepstral feature vector.

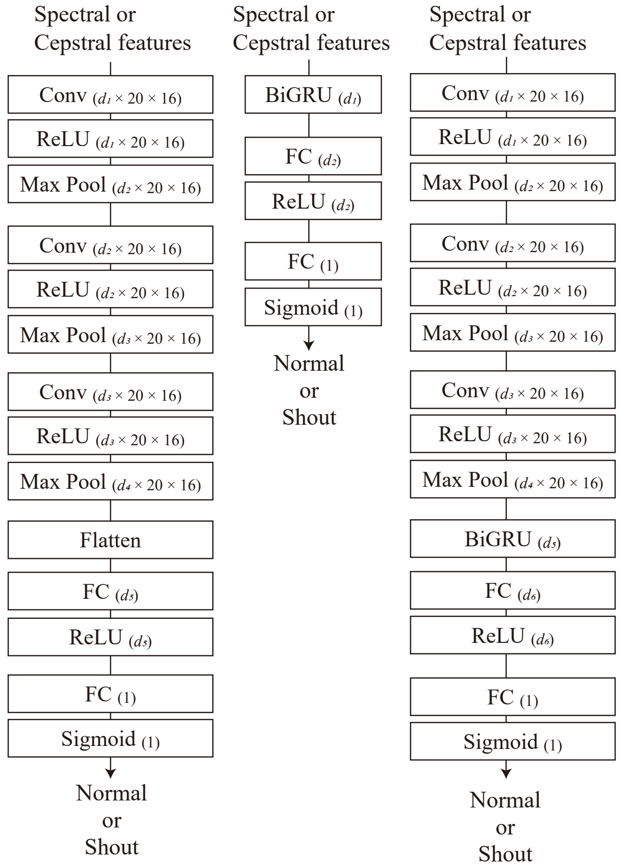
Mel spectrogram: It belongs to spectral features and has been used in recent studies pertaining to sound event detection [3, 20, 28, 29], whose dimensions range between 25 and 40. We extracted a 30-dimensional mel-spectrogram whose number of dimensions is the same as that of the MFCCs.

Herein, we propose learning features that are suitable for shouted and normal speech classification instead of using the conventional feature extractors above. The features used in this study (hereinafter, **our high-level features**) are described below.

Spectrogram: A spectrogram represents the temporal variation of a spectrum. Specifically, applying the short-term Fourier transform to a speech signal yields a 512-dimensional vector of the power spectrum for each frame, and concatenating the vectors of 20 frames results in a 10,240-dimensional spectrogram vector. Recent studies regarding sound event detection used spectrograms as inputs to DNNs and demonstrated their descriptiveness in their target tasks [21–23, 30]. Hence, we used this high-dimensional spectrogram to learn the effective spectral features.

Cepstrogram: Applying the inverse discrete Fourier transform to the log power spectrum yields the cepstrum, and the concatenation of the cepstra of multiple frames yields the cepstrogram. The cepstrogram represents the temporal variation in the vocal tract and vocal cords. We set the dimensionality of each cepstrum to that of the spectrogram, i.e., 512, which resulted in a 10,240-dimensional cepstrogram vector.

The performance of each feature was investigated experimentally.



(a) *Single CNN.* (b) *Single GRU.* (c) *Single CNN-GRU.*

Figure 3: *Types of single feature-based networks, which are compared in experiments. The size of each layer's output is denoted by* $(\cdot)$.

3.2. Network architecture

We used a CNN, GRU, and CNN–GRU to model the acoustic and speech features. We trained these networks as classifiers using single features. Figure 3 shows each network's architecture, which contains hyperparameters depending on the number of feature dimensions. The detailed settings are as follows:

A single CNN comprises three convolutional layers with pooling layers followed by two fully connected (FC) layers, as shown in Fig. 3(a). It regards a set from each feature over 20 frames as an image. All convolutional layers contain a 5×5 kernel with a stride of 1, padding of 2, and 16 channels. The max pooling layers contain a 5×1 kernel for our high-dimensional features and a 3×1 kernel for the conventional low-dimensional features. The layer parameters $(d_1, d_2, d_3, d_4,$ and $d_5)$ in the figure were set as $(512, 102, 20, 4,$ and 64 respectively$)$ for the high-dimensional features and $(30, 10, 3, 1,$ and 16 respectively$)$ for the low-dimensional features.

A single GRU comprises a bidirectional GRU (BiGRU) layer and two FC layers, as shown in Fig. 3(b). Its input is a time series of features from 20 frames. The layer parameters d_1 and d_2 in the figure were set as $(d_1, d_2) = (1024, 64)$ for the high-dimensional input features and $(d_1, d_2) = (60, 16)$ for the low-dimensional ones.

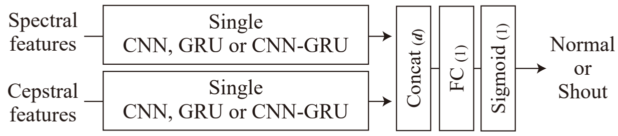


Figure 4: *Spectral-cepstral fusion network. Note that* $(\cdot)$ *indicates the output size for each layer.*

Table 1: *Experimental conditions.*

Training data	Normal speech : 400 samples Shouted speech : 400 samples (female:male = 1:1)
Testing data	Normal speech : 100 samples Shouted speech : 100 samples (female:male = 1:1)
Noise	White noise from NOISEX-92
Sampling	16,000 Hz / 16 bit
SNR	$\infty, -20, -10, -5, 0, 5, 10, 20$ dB
Window length	1,024 points (64 ms)
Hop length	512 points (32 ms)
Window function	Hamming window

A single CNN–GRU comprises three sets of convolutional and pooling layers, followed by a BiGRU layer and two FC layers, as shown in Fig. 3(c). It uses feature images as inputs, and the output of the third max pooling layer is forwarded to the BiGRU as a time series of frame features. We set the parameters of the convolutional and pooling layers (i.e., d_1 to d_5 in the figure) to the same values as those of the single CNN. The remaining parameter d_6 was set to 64 and 16 for the high-dimensional and conventional low-dimensional features, respectively.

Each network uses rectified linear units (ReLUs) as activation functions and a sigmoid function to output the classification result. The mean squared error was used as a loss function to train the network.

3.3. Spectral–cepstral fusion for classification

Finally, the proposed method uses the features from the two domains. Figure 4 shows our DNN architecture comprising the two single networks described in Section 3.2 and an FC layer. First, we pretrained the single feature-based networks using spectral or cepstral features. Subsequently, we concatenated the outputs from the last ReLU layers of these two single networks and input them to the FC layer. The number of dimensions of the concatenated features, d, was 128 for the high-dimensional features and 32 for the low-dimensional ones. The output of the FC layer was forwarded to the sigmoid function to obtain the final classification result. We fine-tuned the entire network using a training dataset, resulting in a feature extractor specific to shouted and normal speech classification.

4. Experiments

4.1. Corpus construction

We first constructed a corpus comprising shouted and normal speech of 40 Japanese speakers (19 females and 21 males). Each speaker was instructed to calmly utter 50 sentences (e.g., "help" and "ahhhhh" in Japanese) and then shout the same sentences. The former and latter words were labeled with "normal"

Table 2: *Comprehensive evaluation of F-measures for different combinations of features and DNN architectures. Spectrogram + Cepstrogram stands for the proposed features and achieved the best performance in all networks.*

Speech features	Model	SNR [dB]								Avg.
		∞	20	10	5	0	-5	-10	-20	
MFCCs_$\Delta\Delta$ [6]	DNN	0.928	0.860	0.833	0.808	0.753	0.683	0.564	0.310	0.717
Mel spectrogram [3]	CNN	0.968	0.950	0.951	0.951	0.943	0.919	0.860	0.726	0.909
tMFCCs [17, 24]		0.968	0.950	0.951	0.951	0.943	0.919	0.860	0.726	0.905
Mel specrogram + tMFCCs [4]		0.968	0.950	0.951	0.951	0.943	0.919	0.860	0.726	0.923
Spectrogram		0.972	0.962	0.963	0.964	0.958	0.943	0.913	0.765	0.930
Cepstrogram		0.961	0.958	0.951	0.946	0.940	0.911	0.895	0.790	0.919
Spectrogram + Cepstrogram		0.977	**0.968**	0.966	**0.967**	0.961	0.947	0.926	0.787	0.938
Mel spectrogram [3]	GRU	0.958	0.944	0.943	0.935	0.913	0.836	0.659	0.401	0.823
tMFCCs [17, 24]		0.972	0.840	0.802	0.794	0.792	0.791	0.791	0.791	0.822
Mel specrogram + tMFCCs [4]		0.981	0.957	0.955	0.950	0.931	0.829	0.789	0.791	0.898
Spectrogram		0.968	0.954	0.957	0.955	0.942	0.902	0.849	0.774	0.913
Cepstrogram		**0.978**	0.938	0.910	0.885	0.847	0.796	0.767	0.579	0.837
Spectrogram + Cepstrogram		0.976	0.961	0.958	0.959	0.948	0.915	0.879	0.720	0.914
Mel spectrogram [3]	CNN-GRU	0.967	0.949	0.947	0.948	0.945	0.909	0.865	0.717	0.906
tMFCCs [17, 24]		0.962	0.948	0.938	0.940	0.926	0.905	0.868	0.745	0.904
Mel specrogram + tMFCCs [4]		0.973	0.960	0.956	0.958	0.951	0.920	0.895	0.758	0.921
Spectrogram		0.976	0.961	0.958	0.949	0.946	0.940	0.918	0.758	0.926
Cepstrogram		0.958	0.940	0.940	0.938	0.934	0.910	0.891	0.765	0.909
Spectrogram + Cepstrogram		**0.978**	0.967	**0.968**	0.960	**0.962**	**0.957**	**0.934**	**0.800**	**0.941**

and "shout," respectively. As manual verification, a graduate student carefully verified all the speech and removed utterances whose categories were considered ambiguous. The remaining number of utterances was 1,116. Finally, the numbers of utterances from the male and female participants were manually set to the same, resulting in 1,000 utterances.

We randomly partitioned our corpus into 800 training-validation and 200 testing speeches. Table 1 summarizes the details of this corpus. As explained in Section 3.1, we partitioned each utterance into successive frames and used a set of 20 frames as the sample to be classified. To consider different noisy conditions in the test, we used NOISE-X92 [31] to add white noise that had the following eight signal-to-noise ratios (SNRs): ∞, 20, 10, 5, 0, -5, -10, and -20 dB.

4.2. Evaluation of classification results

We implemented the networks shown in Figs. 3 and 4 using PyTorch. All the networks were trained using the Adam optimizer with an initial learning rate of 0.001 and momentum parameters of 0.9, and 0.999 on a GeForce GTX 1060 GPU. The batch size was 50, and 100 epochs were used for training. We used the F-measure as a performance evaluation metric.

Table 2 shows the comprehensive evaluation results of different types of features and network architectures in eight SNR conditions, where the averaged F-measures are provided as well. In the table, "tMFCCs" indicates the time series of the MFCCs, and "MFCCs_$\Delta\Delta$" represents the MFCCs and their second derivatives, which yielded the best performance when the state-of-the-art shouted speech detection method was used [6]. The DNN used with MFCCs_$\Delta\Delta$ was the same as that used in [6]. The symbol "+" in the table represents the use of the corresponding two features in a fusion network, as shown in Fig. 4. The performance of the single features shows that our high-level features (i.e., spectrogram or cepstrogram) achieved better F-measures than conventional low-level features in the same domain (i.e., mel spectrogram or MFCCs). Combining the features from the two domains improved the classification performance: the highest F-measures were achieved by "Spectrogram + Cepstrogram." This result shows that our approach can extract features that are suitable for classification from high-dimensional features in different domains.

Comparing the network architectures, the CNNs and CNN–GRUs achieved higher F-measures than the GRUs. This implies that the convolutional layers effectively learned the temporal features.

Finally, we focused on the classification performance under different SNR conditions: in the SNR ranges of -5 dB and ∞ dB, most of the methods achieved F-measures that exceeded 0.9. Meanwhile, in highly noisy environments in which the SNR was -10 dB or -20 dB, the F-measures of the high-level features were higher than those of the low-level features. In particular, only "Spectrogram + Cepstrogram" with the CNN–GRU achieved an F-measure of 0.8 in an environment with an SNR of -20 dB. This shows that our high-level features with the CNN–GRU performed stably and are robust to noise.

5. Conclusions and Future Work

A deep spectral–cepstral fusion approach for shouted and normal speech classification was presented herein. We investigated various DNN architectures in the proposed framework. A comprehensive comparison revealed that learning features from both spectral and cepstral domains facilitated the target task effectively. In particular, the CNN–GRU architecture with high-dimensional features yielded the best classification performance in various noisy environments.

In the future, we plan to construct a larger, more varied corpus comprising shouts and normal speech and release it online. Additionally, we will extend the current binary classification of speech status to the multiclass classification of intensity.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K14381.

7. References

[1] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," *Sensors*, vol. 18, no. 6, 2018.

[2] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58 043–58 055, 2018.

[3] M. Valenti, D. Tonelli, F. Vesperini, E. Principi, and S. Squartini, "A neural network approach for sound event detection in real life audio," in *European Signal Processing Conference (EUSIPCO)*, 2017, pp. 2754–2758.

[4] J. F. Gaviria, A. Escalante-Perez, J. C. Castiblanco, N. Vergara, V. Parra-Garces, J. D. Serrano, A. F. Zambrano, and L. F. Giraldo, "Deep learning-based portable device for audio distress signal recognition in urban areas," *Applied Sciences*, vol. 10, no. 21, 2020.

[5] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, "Audio-based event detection at different snr settings using two-dimensional spectrogram magnitude representations," *Electronics*, vol. 9, no. 10, 2020.

[6] S. Baghel, S. R. M. Prasanna, and P. Guha, "Exploration of excitation source information for shouted and normal speech classification," *The Journal of the Acoustical Society of America*, vol. 147, no. 2, pp. 1250–1261, 2020.

[7] L. Bailly, N. H. Bernardoni, F. Müller, A.-K. Rohlfs, and M. Hess, "Ventricular-fold dynamics in human phonation," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1219–1242, 2014.

[8] A. Anikin, R. Bååth, and T. Persson, "Human non-linguistic vocal repertoire: Call types and their meaning," *Journal of Nonverbal Behavior*, vol. 42, pp. 53–80, 2018.

[9] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[10] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics", 2014, pp. 103–111.

[11] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4968–4971.

[12] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," in *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[13] A. Sharma and S. Kaul, "Two-stage supervised learning-based method to detect screams and cries in urban environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 290–299, 2016.

[14] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *IEEE Conference on Industrial Electronics and Applications*, 2010, pp. 2115–2120.

[15] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 165–168.

[16] H. Nanjo, H. Mikami, S. Kunimatsu, H. Kawano, and T. Nishiura, "A fundamental study of novel speech interface for computer games," in *IEEE International Symposium on Consumer Electronics*, 2009, pp. 558–560.

[17] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6460–6464.

[18] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[19] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5089–5093.

[20] S. Lee, M. Kim, S. Shin, S. Park, and Y. Jeong, "Data-dependent feature extraction method based on non-negative matrix factorization for weakly supervised domestic sound event detection," *Applied Sciences*, vol. 11, no. 3, 2021.

[21] Y. Zhang, K. Zhang, J. Wang, and Y. Su, "Robust acoustic event recognition using AVMD-PWVD time-frequency image," *Applied Acoustics*, vol. 178, p. 107970, 2021.

[22] G. Ciaburro, "Sound event detection in underground parking garage using convolutional neural network," *Big Data and Cognitive Computing*, vol. 4, no. 3, 2020.

[23] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing*, vol. 272, pp. 505–512, 2018.

[24] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "Deep neural network based learning and transferring mid-level audio features for acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 796–800.

[25] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 161–165.

[26] J. L. Rouas, J. Louradour, and S. Ambellouis, "Audio events detection in public transport vehicle," in *IEEE Intelligent Transportation Systems Conference*, 2006, pp. 733–738.

[27] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 158–161.

[28] A. Vafeiadis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Audio content analysis for unobtrusive event detection in smart homes," *Engineering Applications of Artificial Intelligence*, vol. 89, no. 103226, 2020.

[29] J. Schröder, J. Anemüller, and S. Goetze, "Performance comparison of GMM, HMM and DNN based approaches for acoustic event detection within task 3 of the DCASE 2016 challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, pp. 80–84.

[30] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 559–563.

[31] "NOISE-X92," http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html, Last accessed: 22/03/2021.