IEICE TRANS. FUNDAMENTALS, VOL.E107–A, NO.4 APRIL 2024

LETTER *Special Section on Information and Communication Technologies for Safe and Secure Life*

Noise-Robust Scream Detection Using Wave-U-Net

Noboru HAYASAKA[†a)], ***Member***, **Riku KASAI**[††], ***Nonmember***, ***and*** **Takuya FUTAGAMI**[†], ***Member***

SUMMARY In this paper, we propose a noise-robust scream detection method with the aim of expanding the scream detection system, a sound-based security system. The proposed method uses enhanced screams using Wave-U-Net, which was effective as a noise reduction method for noisy screams. However, the enhanced screams showed different frequency components from clean screams and erroneously emphasized frequency components similar to scream in noise. Therefore, Wave-U-Net was applied even in the process of training Gaussian mixture models, which are discriminators. We conducted detection experiments using the proposed method in various noise environments and determined that the false acceptance rate was reduced by an average of 2.1% or more compared with the conventional method.

key words: *scream detection, scream enhancement, Wave-U-Net*

1. Introduction

Security cameras have been installed in various places to prevent and deter crimes. However, installation locations are limited due to privacy or brightness requirements. To address these disadvantages, security systems that use sound have been proposed. Scream detection systems can detect abnormal conditions immediately, so it is highly effective in preventing and deterring crimes [1]–[6]. In addition, the recorded screams may be used as evidence in trials and investigations [7].

In this paper, we define a scream as a sound made by a woman to express fear. The reason is that women are generally less self-protective than men and are more likely to scream. The scream detection system has the advantage of being usable anywhere, but it can be affected by noise and may not fully serve its purpose. In addition, if the recorded screams are degraded by noise, they lose their validity as evidence. Therefore, in order to preserve the evidential value, we conducted comparative experiments on noise reduction methods for noisy screams. The results verified that Wave-U-Net reduces the most noise compared with the other methods, e.g. Speech Enhancement Generative Adversarial Network [8]. However, we also found that Wave-U-Net emphasizes noise components similar to screams, which lowers its accuracy [9].

Manuscript received March 7, 2023.
Manuscript revised July 13, 2023.
Manuscript publicized October 5, 2023.

[†]The authors are with the Department of Engineering Informatics, Osaka Electro-Communication University, Neyagawa-shi, 572-8530 Japan.
[††]The author is with OPTAGE Inc., Osaka-shi, 540-8622 Japan.
a) E-mail: hayasaka@osakac.ac.jp

DOI: 10.1587/transfun.2023SSL0001

Audio surveillance systems such as sound event detection have been studied, but few studies have specialized in scream detection systems. Mel-frequency cepstral coefficients (MFCCs), band-limited spectral entropy [2], and Combo-SAD, which integrates time domain features and frequency domain features [3], have been proposed as features of scream detection. The Gaussian mixed model (GMM) and support vector machine (SVM) are widely used classifiers, and a method of tuning SVM parameters according to signal-to-noise ratio (SNR) and the context of input audio samples has also been proposed [4]. Moreover, several methods using deep learning have also been proposed [5]–[6]. In this paper, we investigate whether it is possible to improve the accuracy of scream detection by using enhanced screams with Wave-U-Net. Therefore, we use MFCCs and GMMs, which are widely used in conventional scream detection, as the features and classifiers. The use of deep learning for both scream enhancement and scream detection increases computational costs and is not discussed in this paper.

Section 2 explains Wave-U-Net and the enhanced screams. The framework for scream detection using Wave-U-Net is described in Sect. 3, and the evaluation results are presented in Sect. 4. Finally, the key points are summarized in Sect. 5.

2. Wave-U-Net for Scream Enhancement

2.1 Wave-U-Net [10]

The Wave-U-Net architecture is a one-dimensional version of the general u-net that can directly handle time domain signals. Wave-U-Net is used for separating music and vocals. We have also found that it is highly effective for separating screams and noise [9].

Figure 1 shows the architecture of Wave-U-Net. It contains L downsampling blocks, which each consist of a one-dimensional convolution and decimation layer, one bottom convolution layer, and L upsampling blocks, which each consists of a one-dimensional convolution and interpolation layer. The input signals are noisy screams and the output signals are the clean screams and noise.

The downsampling blocks extract a number of higher-level features while reducing the time resolution. These features are concatenated with local high-resolution features calculated from the same level upsampling blocks. The results are concatenated into multi-scale features for prediction. The decimation layer in each downsampling block

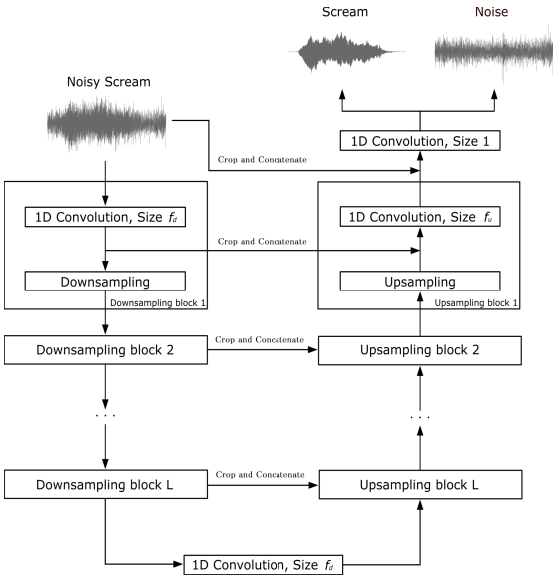

Fig. 1 Wave-U-Net architecture.

operates with half the time resolution of the previous block. The one-dimensional convolution layer in a downsampling block has $F * l$ filters of size f_d, where l denotes the order of the downsampling blocks.

Each upsampling block executes double upsampling in the time direction, followed by concatenating features from the same-scale downsampling blocks and then one-dimensional convolution. Bilinear interpolation is used in each interpolation layer. The one-dimensional convolution layer in an upsampling block has $F * l$ filters of size f_u.

Each convolution layer in these blocks is followed by leaky rectified linear unit activation with $\alpha = 0.3$, and tanh is used in the last convolution layer of the network.

2.2 Enhanced Scream with Wave-U-Net

Figure 2 shows the spectrograms of a clean scream, noisy scream, and enhanced scream with Wave-U-Net. In these spectrograms, the screaming section is between 0.25 s and 1.25 s. Wave-U-Net succeeded in largely removing noise and leaving the harmonic components. However, in the non-screaming section (0 to 0.1 s and 1.3 to 1.5 s) in Fig. 2(c), the same frequency component as the scream was enhanced. In addition, the spectrograms of the clean scream and the emphasized scream are different in the screaming section. Therefore, if the output from Wave-U-Net is used for scream detection, mis-detection may frequently occur in the non-screaming section.

3. Noise-Robust Scream Detection

3.1 Scream Detection Framework Using Wave-U-Net

The proposed scream detection framework is shown in Fig. 3. The highlight of this framework is that Wave-U-Net is also

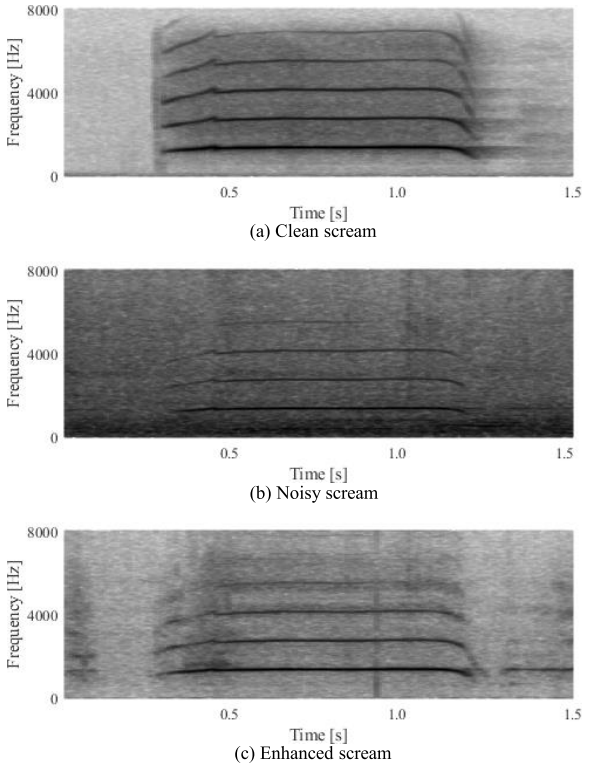

Fig. 2 Spectrograms of different screams.

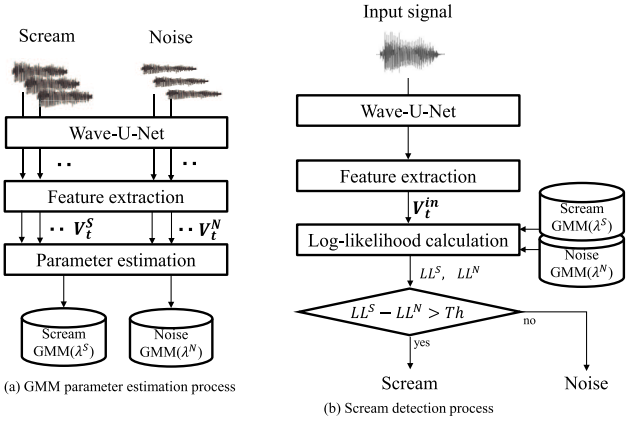

Fig. 3 Framework of scream detection using Wave-U-Net.

applied in the parameter estimation process to solve the problems described in Sect. 2.2. In the feature extraction step of Fig. 3, the MFCCs described in the next section are extracted.

Here, we describe the parameter estimation process in Fig. 3(a). $\boldsymbol{V}_t^S$ and $\boldsymbol{V}_t^N$ are the MFCCs of the scream and the noise, respectively, and t is the frame number. In the parameter estimation step, the MFCCs are modeled using GMMs (λ^{S} and λ^{N}).

In the detection process shown in Fig. 3(b), $\boldsymbol{V}_t^{\mathrm{in}}$ is derived from the input signal, and the log likelihoods for the respective GMMs are calculated: (LL_t^{S} and LL_t^{N}).

$$LL_t^{\mathrm{S}} = \log p(\boldsymbol{V}_t^{\mathbf{in}}|\lambda^{\mathrm{S}}) \quad (1)$$

$$LL_t^{\mathrm{N}} = \log p(\boldsymbol{V}_t^{\mathbf{in}}|\lambda^{\mathrm{N}}) \quad (2)$$

When the difference (LL_t^{S} - LL_t^{N}) exceeds the threshold (Th), the input signal is judged to be a scream. The optimal value of Th depends on how many undetected screams and misdetected noise can be tolerated. If the environment in which this system is used can be predicted, it is desirable to determine Th experimentally from the environmental noise and the screams used for training. On the other hand, if it cannot be predicted, it is necessary to determine Th experimentally from the noise and screams used for training.

3.2 Mel-Frequency Cepstral Coefficients

Because the fundamental frequency and the log-energy, which are prosodic features, significantly deteriorate due to noise, MFCCs are used as phonemic features instead of the prosodic features.

MFCCs, which are cepstral coefficients that take human hearing characteristics into account, are used as feature vectors representing the vocal tract. They are also widely used in speech recognition, speaker recognition, and other related tasks. The l^{th} MFCC ($C_t[l]$) is calculated using the following equations.

$$C_t[l] = \sqrt{\frac{1}{M}} \sum_{m=0}^{M-1} \log\left(X_t^{\mathrm{mel}}[m]\right) \cos\left(\frac{(2m+1)\pi l}{M}\right) \quad (3)$$

$$X_t^{\mathrm{mel}}[m] = \sum_{k=0}^{K-1} B_{m,k} |X_t[k]|^2 \quad (4)$$

The $B_{m,k}$ is the mel-filterbank matrix used in the ETSI standard front-end [11], m and k are the filter bank number and frequency bin, respectively, $X_t[k]$ is a spectrum, and M is the number of filter banks. The value of l is taken as $1 \leq l \leq 12$.

4. Experiments

4.1 Set-Up

We used the screams from 40 people in the scream database described in [2]. The total number of screams was 705, for a total duration of 1400 s. The screams were divided into two sets of 20 people each, one for training Wave-U-Net and the scream GMM, and the other for testing. The number of screams was 438 for training and 267 for testing. The data were downsampled to 16 kHz because it has been shown that the main component of screams exists below 8 kHz [2].

Six types of noise data ('station', 'factory', 'intersection', 'train', 'computer room', 'air conditioner') were selected from the Japan Electronic Industry Development Association (JEIDA) noise database [12]. To compare the performance of known and unknown noises, 'station', 'factory', and 'intersection' were designated as the known noise set, and 'train', 'computer room', and 'air conditioner' were designated as the unknown noise set. The known noise set was used for training Wave-U-Net and the noise GMM. The number of noise frames was 454,240 for training and 451,842 for testing. Noisy screams for testing were superimposed on the scream of the testing set with SNR = 0 dB.

Table 1 Analysis conditions.

Window function	Hanning window
Frame length	512 samples
Shift length	256 samples
Mel-filterbank outputs	24

Wave-U-Net models were trained on randomly sampled audio excerpts using the Adam optimizer (learning rate=0.0001, decay rates β_1=0.9, and β_2=0.999) with a batch size of 16. Following a previous study [10], our network layer size was 12, and we set $F = 24$ extra filters for each layer with downsampling block filters of size $f_d = 15$ and upsampling block filters of size $f_u = 5$.

Feature extraction was performed with the analysis conditions listed in Table 1, and the number of mixtures in the GMMs was fixed at 32. We determined the initial values of all GMMs by the k-means method. In the conventional method, we did not emphasize screams using Wave-U-Net. The proposed and conventional methods were evaluated with the performance measure FAR_{min}.

$$FAR[\%] = \frac{\text{Num. of misdetected noise frames}}{\text{Number of evaluated noise frames}} \times 100 \quad (5)$$

$$FRR[\%] = \left(1 - \frac{\text{Number of detected screams}}{\text{Number of evaluated screams}}\right) \times 100 \quad (6)$$

$$FAR_{\mathrm{min}} = \min\ FAR, \quad \text{subject to}\ \ FRR = 0 \quad (7)$$

Here, FAR and FRR represent False Acceptance Rate and False Rejection Rate, respectively. Considering the purpose of the scream detection system, it is necessary to detect all screams. Therefore, FAR_{min} was used for evaluation. The experiments compare the following four methods.

- Method 1: Do not apply Wave-U-Net to parameter estimation or detection (conventional method).
- Method 2: Apply Wave-U-Net to detection, but not to parameter estimation.
- Method 3: Apply Wave-U-Net to the detection and parameter estimation of the scream GMM, but not to the parameter estimation of the noise GMM.
- Method 4: Apply Wave-U-Net to parameter estimation and detection (proposed method).

4.2 Results and Discussions

The experimental results are shown in Table 2. Compared with Method 1, which is the conventional method, Method 2 detected screams more accurately, indicating that the emphasized scream is effective for detection. Next, between Method 2 and Method 3, Method 3 was slightly more accurate. From this, it can be said that Wave-U-Net should be applied even when estimating the parameters of the scream GMM because the frequency characteristics of clean screams and enhanced screams are different. Finally, Method 4, the proposed method, was the most effective in most noisy environments, with an average improvement of about 2.1%

Table 2 Experimental results [%].

Types of noise	Method1	Method2	Method3	Method4
Station	0.447	0.201	0.145	**0.021**
Factory	0.276	0.065	0.016	**0.015**
Intersection	0.271	0.024	0.014	**0.001**
Train	11.583	0.932	0.989	**0.544**
Computer room	0.016	**0.000**	**0.000**	**0.000**
Air conditioner	1.035	0.003	**0.002**	0.007
Average	2.271	0.204	0.194	**0.098**

compared with the conventional method. Thus, in scream detection using Wave-U-Net, the optimal detection can be obtained by applying Wave-U-Net even when estimating the parameters of GMMs.

When GMMs are used as discriminators, the detection performance depends on its initial values and discrimination threshold (Th). In particular, Th should be determined carefully as it depends on usage conditions. Although the computational cost increases, it is necessary to consider threshold-independent discriminators using deep learning in the future.

5. Conclusion

In this paper, we proposed a noise-robust scream detection method using enhanced screams with Wave-U-Net. The enhanced screams show different frequency characteristics from those of clean screams because the harmonic components deteriorate. Therefore, Wave-U-Net was purposely applied to the clean screams to train the scream GMM. The results of the scream detection experiments showed that the $FAR_{\min}$ could be reduced by 2.1% compared with the conventional method. In the future, we aim to simplify the network structure of Wave-U-Net and develop mobile applications.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 19K04935.

References

[1] J.T. Geiger and K. Helwani, "Improving event detection for audio surveillance using Gabor filterbank features," European Signal Processing Conference (EUSIPCO), pp.719–723, 2015.
[2] N. Hayasaka, A. Kawamura, and N. Sasaoka, "Noise-robust scream detection using band-limited spectral entropy," AEU-International Journal of Electronics and Communications, vol.76, pp.117–124, 2017.
[3] M.K. Nandwana, A. Ziaei, and J.H.L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.161–165, 2016.
[4] A. Sharma and S. Kaul, "Two-stage supervised learning-based method to detect screams and cries in urban environments," IEEE/ACM Trans. Audio, Speech, Language Process., vol.24, no.2, pp.290–299, 2015.
[5] P. Laffitte, D. Sodoyer, C. Tatkeu, and L. Girin, "Deep neural networks for automatic detection of screams and shouted speech in subway trains," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6460–6464, 2016.
[6] T. Fukumori, "Deep spectral-cepstral fusion for shouted and normal speech classification," Interspeech 2021, pp.4174–4178, Sept. 2021.
[7] R.C. Maher, Principles of Forensic Audio Analysis, Chapter 6, Springer, Switzerland, 2018.
[8] S. Pascual, A. Bonafonte, and J. Serrá, "SEGAN: Speech enhancement generative adversarial network," arXiv:1703.09452, 2017.
[9] R. Kasai, N. Hayasaka, T. Futagami, and Y. Miyanaga, "Scream enhancement using Wave-U-Net," International Workshop on Smart Info-Media Systems in Asia (SISA), pp.5–8, Sept. 2021.
[10] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," Proc. 19th Int'l Society for Music Information Retrieval Conference (ISMIR), Sept. 2018.
[11] Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms, European Telecommunications Standards Institute 201 108 V1.1.3, Sept. 2003.
[12] JEIDA Noise Database (ELRA-SD37), http://universal.elra.info/product_info.php?cPath=37_39&products_id=53