Int. J. Electron. Commun. (AEÜ) 76 (2017) 117–124

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

journal homepage: www.elsevier.com/locate/aeue

Regular paper

Noise-robust scream detection using band-limited spectral entropy

Noboru Hayasaka [a,*], Arata Kawamura [b], Naoto Sasaoka [c]

[a] Osaka Electro-Communication University, 18-8 Hatsucho, Neyagawa-shi, Osaka, Japan
[b] Osaka University, 1-3 Machikaneyamacho, Toyonaka-shi, Osaka, Japan
[c] Tottori University, 4-101 Koyamacho Minami, Tottori-shi, Tottori, Japan

ARTICLE INFO

Article history:
Received 15 June 2016
Accepted 3 April 2017

Keywords:
Scream detection
Spectral entropy
Band-limited

ABSTRACT

This report describes scream detection systems that can detect screams under noisy conditions and describes techniques for increasing their noise-robustness. More specifically, spectral entropy, which expresses the difference in frequency distribution between a scream and noise, is used as a detection feature. Furthermore, a method is presented that improves scream detection accuracy in noisy environments by limiting the frequency band of the spectral entropy used as the detection feature. Evaluation experiments in noisy environments demonstrated that the proposed method has better scream detection capability than a conventional method. The proposed method can detect screams with equal error rates of 0.3% and 0.8% under 0 and −5 dB conditions, respectively. Furthermore, the cost of likelihood calculation is about 1/12th that of the conventional method. The proposed method can thus be used to develop a scream detection system that is sufficiently accurate in an actual environment.
© 2017 Elsevier GmbH. All rights reserved.

1. Introduction

The increasing diversity and sophistication of crime in recent years has increased the demand for security systems. Video cameras are typically used in security systems to capture an image of a crime scene and the appearances of the individuals involved. In addition to capturing evidence of a crime, they help prevent crime if they are placed in conspicuous locations. However, cameras pose privacy problems and thus cannot be placed in such locations as restrooms and changing rooms. Even in locations where they can be placed, there are various other problems, such as the need for special cameras to deal with darkness.

To address these disadvantages, several security systems that use sound (such as a scream) have been proposed [1–8]. Most of them use a classifier: a Gaussian mixture model (GMM) [1,2], a support vector machine (SVM) [3–5], a hidden Markov model (HMM) [6,7], a deep neural network (DNN) [8], and so on [9]. Prosodic features [4], spectral features [1,5,7,8], the zero-crossing rate [1,7], and mel-frequency cepstral coefficients (MFCCs) [1–6] are conventionally used in regard to fea tures. In this paper, a method is presented that uses band-limited spectral entropy to improve detection performance and a GMM to classify the extracted feature. A GMM is used because it is easy to expand into multi-class discrimination. Screams, which are defined here as a sound made by female humans to convey fear, are generally characterized by a high fundamental frequency. The spectral entropy feature represents the whiteness of a spectrum and can be easily distinguished between screams and other sounds. Furthermore, since spectral entropy can be used to evaluate the shape of the spectral distribution without the effect of the magnitude, spectral entropy has high noise-robustness.

The novelty of this study is the development of a system which can detect screams only with high noise-robustness. The systems in [4,9] have dealt with screams only, but the most of the conventional audio security systems have dealt with various sounds including a scream. Moreover, the systems in [4,9] have used prosodic information and/or MFCCs. The novelty of this study is also to propose a feature specialized for detection of scream under noisy environments.

Section 2 explains the features and framework that are conventionally used for scream detection and describes the low noise-robustness of methods that use them. The proposed method using spectral entropy is described in Section 3, and the evaluation results are presented in Section 4. Finally, the key points are summarized in Section 5.

2. Conventional features and scream detection framework

2.1. Scream detection using prosodic features

Since screams have both high energy and a high fundamental frequency, the prosodic features of the log-energy and fundamental frequency have been used for scream detection [4]. Because the

* Corresponding author.
E-mail addresses: hayasaka@osakac.ac.jp (N. Hayasaka), kawamura@sys.es.osaka-u.ac.jp (A. Kawamura), sasaoka@ele.tottori-u.ac.jp (N. Sasaoka).
URL: http://www.osakac.ac.jp/ (N. Hayasaka).

http://dx.doi.org/10.1016/j.aeue.2017.04.002
1434-8411/© 2017 Elsevier GmbH. All rights reserved.

prosodic features depend on the environment and the speaker, they are directly processed using thresholding (Fig. 1) rather than a GMM as used in the proposed method and other statistical methods.

The log-energy (E_t) and fundamental frequency (FF_t) at the tth frame are extracted using the equations in Appendix A. The E_t and FF_t of screams in environments with SNRs of 20, 10, and 0 dB are shown in Fig. 2. The extraction conditions are given in Section 4.2. The superimposed noise was a train sound from the JEIDA NOISE database [10]. In Fig. 2, there is a scream signal in frames 280 through 380, when the noise condition was that of a train gradually approaching. The difference in E_t between the scream and non-scream parts decreased, particularly for the 0 dB environment, for which the scream part cannot be distinguished from the non-scream part (after frame 600). Additionally, although the scream had an FF_t of 800–1300 Hz, the FF_t was incorrectly extracted for the scream part, and the same FF_t was extracted for the non-scream part for the 0 dB environment. Thus, prosodic features are susceptible to the effects of noise and unsuitable as features for scream detection under noisy conditions.

2.2. Scream detection using phonemic features

The susceptibility of prosodic features to degradation due to noise, as described above, led researchers to use MFCCs as phonemic features in addition to or instead of prosodic features. MFCCs, which are cepstral coefficients that take human hearing characteristics into account, are used as a feature vector representing differences in the configuration of the vocal tract. They are widely used in speech recognition, speaker recognition, and similar areas. A scream conveys fear through the use of a common vocal tract configurationthe mouth is opened as if to pronounce ‘a’ as a phonemic suffix. This suggests that MFCCs are effective for scream detection. The lth MFCC ($C_t[l]$) is calculated using the following equations.

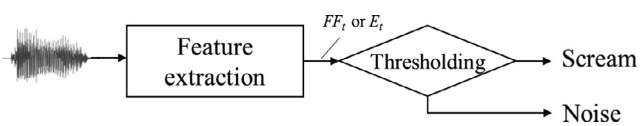

Fig. 1. Framework of scream detection using prosodic features.

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

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

The $B_{m,k}$ is the mel-filterbank matrix used in the ETSI standard front-end [11], m is the filter bank number, $X_t[k]$ is the spectrum represented by Eq. (A.2), and M is the number of filter banks. For scream detection, as described in Huang et al. [4], and speech recognition, the value of l is taken as $1 \leqslant l \leqslant 12$.

2.3. Framework of scream detection using GMMs

Framework of scream detection using GMMs is shown in Fig. 3. The features of the scream and the noise (non-scream sounds), $\boldsymbol{V_t}$, are modeled using GMMs represented by the following probability density functions (Fig. 3(a)).

$$p(\boldsymbol{V_t}|\lambda^{\mathrm{S}}) = \sum_{d=1}^{D^{\mathrm{S}}} w_d^{\mathrm{S}} \mathcal{N}(\boldsymbol{V_t}|\boldsymbol{\mu_d^{\mathrm{S}}}, \boldsymbol{\Sigma_d^{\mathrm{S}}}) \quad (3)$$

$$\lambda^{\mathrm{S}} = \{\boldsymbol{\mu_1^{\mathrm{S}}}, \cdots, \boldsymbol{\mu_{D^{\mathrm{S}}}^{\mathrm{S}}}, \boldsymbol{\Sigma_1^{\mathrm{S}}}, \cdots, \boldsymbol{\Sigma_{D^{\mathrm{S}}}^{\mathrm{S}}}, w_1^{\mathrm{S}}, \cdots, w_{D^{\mathrm{S}}}^{\mathrm{S}}\} \quad (4)$$

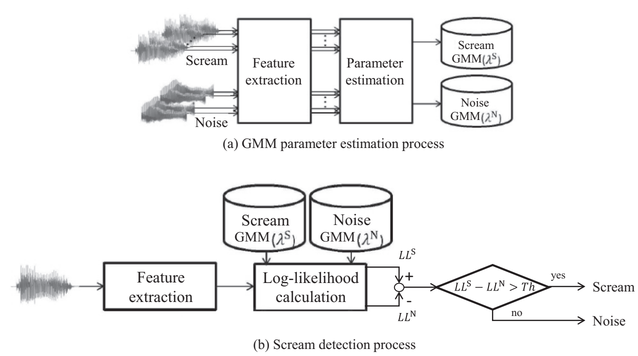

Fig. 3. Framework of scream detection using GMMs.

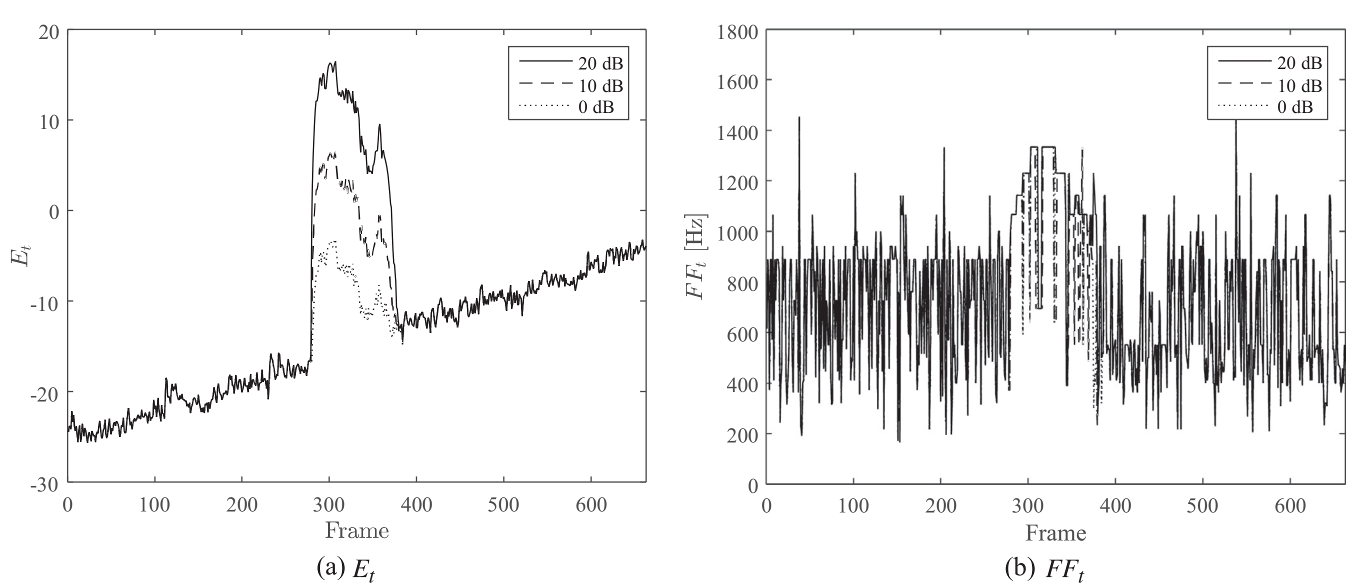

Fig. 2. Frame-time trajectories of prosodic features under noisy conditions.

Download English Version:

https://daneshyari.com/en/article/4954057

Download Persian Version:

https://daneshyari.com/article/4954057

Daneshyari.com