engineering proceedings

Proceeding Paper

Finding Earthquake Victims by Voice Detection Techniques †

Ruchi Jha *, Walter Lang and Reiner Jedermann

Institute for Microsensors, Actuators, and Systems (IMSAS), University of Bremen, 28359 Bremen, Germany; WLang@imsas.uni-bremen.de (W.L.); rjedermann@imsas.uni-bremen.de (R.J.)

* Correspondence: rjha@imsas.uni-bremen.de

† Presented at the 8th International Electronic Conference on Sensors and Applications, 1–15 November 2021; Available online: https://ecsa-8.sciforum.net.

Abstract: After an earthquake or a building collapse, victim recovery is a challenging task. Recovery methods must include the location of victims by non-visual means; human speech is one such parameter that can be used in victim and rescue operations. In this paper, we discuss the application of a voice-detection technique for the discrimination of voice and non-voice sounds, based on frequency parameters such as flux, centroid, and roll-off of audio signals. Using the cross-validation tests based on a linear discriminant analysis model, flux and centroid individually displayed the highest success rates for all categories of test sample. By combining these two parameters, the recognition rate was improved to 78% for the signals with high background noise.

Keywords: voice activity detection; rescue method; noise separation

1. Introduction

A large portion of the world's population is affected by earthquakes every year, and more than 430,000 people have died in earthquakes during the 21st century alone [1]. Most of the earthquakes above 5.5 on the Richter scale can cause large-scale destruction through building collapse and structural damage [2]. Approximately 80 percent of victims can be successfully rescued alive if they are detected by help teams within 48 h [3]. This means that detecting an injured victim and providing medical care in the shortest time is a priority of any disaster rescue operation.

Currently cameras, drones, sensitive microphones, mobile video cameras, and specially trained dogs are used to locate stuck victims [4]. However, rescue is challenging when the victim cannot be found through direct line of sight. An advanced device such as FINDER (Finding Individuals for Disaster and Emergency Response), a product made by NASA, is able to detect a human stuck beneath 30 feet of debris [5]. It employs an advanced system that sends and receives a low-power microwave signal at a disaster site and has the ability to differentiate between human, animal, and mechanical movements. Unfortunately, FINDER is not available commercially, and it is expensive to arrange for its use by local teams.

Previous tests with a thermal camera (requiring line-of-sight), radar-based motion sensor, and a CO_2 gas sensor could not provide a sufficiently high recognition rate [6]. To further enhance the system's performance overall, speech detection methods were investigated. One method of discerning voices from noises or non-human sounds is commonly termed voice activity detection (VAD). A VAD algorithm is usually designed to extract specific features from an input signal, e.g., energy, zero crossing rate, periodicity measure, spectral features—alone or in combination. This is commonly used in speech communication systems such as hands-free telephony, echo cancellation, and speech coding and recognition [7,8]. This paper discusses the application and testing of the VAD algorithm to discriminate speech from non-speech signals.

check for updates

Citation: Jha, R.; Lang, W.; Jedermann, R. Finding Earthquake Victims by Voice Detection Techniques. *Eng. Proc.* **2021**, *10*, 69. https://doi.org/10.3390/ecsa-8-11248

Academic Editor: Stefano Mariani

Published: 1 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2. Methodology

Every signal has a variety of attributes, and those attributes can be broadly categorized into time and frequency parameters [9].

2.1. Frequency Domain Parameters

The following three frequency parameters were selected for current research:

Spectral Flux measures the spectral change between the previous frames of signal to the current frame and expresses how quickly power spectrum of a signal is changing [10]. It can be calculated using the following formula:

$$f(t_n) = \sum_{k=1}^{l} (e_n(k) - e_{n-1}(k))^2 \tag{1}$$

where n and $n-1$ are consecutive windows of length l, and $e_n(k)$ is the kth normalized DFT (discrete Fourier transform) coefficient of the nth frame.

Spectral Centroid (SC) is a measure of the centre of mass of the power spectrum. Higher values of the spectrum centroid suggest brighter sound [11]. For a spectral frame, a centroid is calculated by the mean bin of the power spectrum as follows:

$$SC = \frac{\sum_{n=1}^{l} k\, F(k)}{\sum_{n=1}^{l} F(k)} \tag{2}$$

where $F(k)$ is the amplitude corresponding to bin n in the DFT spectrum.

Spectral roll-off denotes the value of the frequency, below which 95% of signal energy resides. It is the measure of skewness of the shape of the power spectrum and can be used to distinguish signals [8]. f_R is given by the solution of Equation (3).

$$\sum_{n=1}^{f_r} F(k) = 0.95 * \sum_{n=1}^{l} F(k) \tag{3}$$

2.2. Noise and Voice Samples

Using a TIE StudioDynamic Mic, various speech samples from different age groups speaking in different languages were recorded along with some standard recorded noise available via commercial audio CDs [12]. To maintain uniformity, all the recordings were taken at a sampling frequency of 44,100 Hz (mono) using Audacity digital audio software. All the available sound samples were categorized as per Table 1 and further used for analysis.

Table 1. Training dataset.

Group	Sources	Name	Examples
Noise Studio	Audio CD [11]	N1 to 11	Traffic, touring cars, motorcars, cleaning, airplane, buzzer, river, applause, industry, chattering.
Voice Samples	CD, TV, studio recording	VF1 to 4 (female) VM1 to 5 (male)	Female and Male sound recordings in English and German.
Noise Street	Outside recording	SN 1 to 7	Street noises with birds, cars, tram, glasses, music, river and wind
Voice Mix	Outside recording	MIX 1 to 5	Mix sounds of people speaking with background noise.
Voice Studio	Studio recording	VF . . . (female) VM . . . (male)	Speech recorded in Spanish (S), German (D), Hindi (H), English (E), and Latvian (L)

2.3. Post-Processing of Frequency Domain Parameters

The frequency-domain parameters were calculated for short frames of 2 ms, resulting in a time-dependent curve for each parameter. A distinction between voice and noise was not possible based on the simple average values. Keeping this under consideration, the entire sample was searched for peaks values in both positive and negative direction for each of the spectral parameter. The averages of these peaks were calculated as the $\mathrm{P}_{\mathrm{av}+}$ and $\mathrm{P}_{\mathrm{av}-}$ values (example: Figure 1). In this way, the time graphs were compressed to only two parameters, reducing the risk of errors.

2.4. Training

The noise studio and voice samples groups were used for training. The $\mathrm{P}_{\mathrm{av}+}$ and $\mathrm{P}_{\mathrm{av}-}$ values for each training sound sample were marked in the related graphs. A separation line between voice and noise samples was defined based on linear discriminant analysis classification for each graph [13].

2.5. Cross-Validation

The accuracy of our detection system was verified by cross-validation [14]: Samples from so far unused audio samples were classified based on the separation line obtained from the training samples. The share of correctly and falsely classified samples was calculated. The Statistics and Machine Learning Toolbox in MATLAB was used to perform training and cross validations, and DSP Toolbox was used for the audio signal pre-processing.

3. Results

3.1. Peak Detection

Figure 1 shows an example of post-processing frequency domain parameters. The flux values of a rain sound signal were plotted, the positive and negative peak values were automatically marked, and their average was calculated.

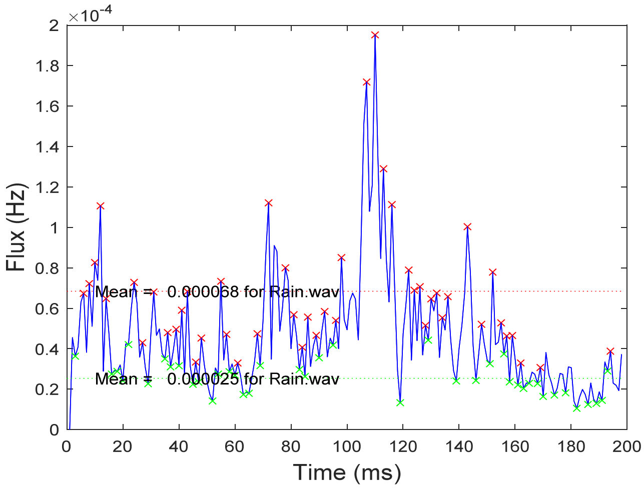

Figure 1. Calculation of $\mathrm{P}_{\mathrm{av}+}$ and $\mathrm{P}_{\mathrm{av}\text{-}}$ for Flux in a rain sound audio sample (red = positive peak values, green = negative peak values).

Similar graphs were plotted for centroid and roll-off values by considering their peak average positive and negative values on the entire training sample. Both the training and testing process is demonstrated in Figure 2. The training audio samples, which were already known to the systems, are shown in blue and red. The green line indicates the trained separation based on linear discriminate analysis.

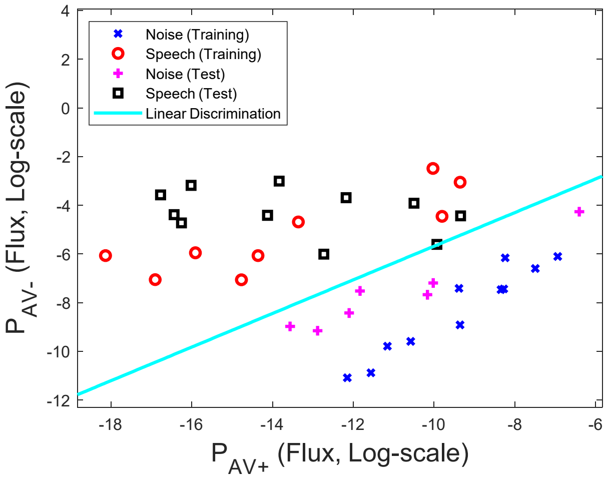

Figure 2. A distinction of noise and speech based on P_{av+} and P_{av-} for flux (training and testing noise in blue and pink; speech in red and black, respectively).

3.2. Cross Validation Results

All the samples, irrespective of their audio group, were correctly distinguished using their P_{av+} and P_{av-} of flux by using the trained linear boundary. Testing the parameters individually based on the automated analysis model of classification a varying success rate from 78% (in case of roll off) to 100% (for flux) was obtained. The ratio of the number of correctly detected sample to the falsely detected samples was used to determine the success rate in Table 2.

Table 2. Results of cross-validation: number of correctly and falsely placed samples and total success rate for cross validation.

Group	Flux	Roll-Off	Centroid	Flux and Centroid	Centroid and Roll Off
Noise Street	7/0	7/0	7/0	7/0	7/0
Voice Mix	5/0	2/3	3/2	5/0	4/1
Voice Studio	6/0	5/1	5/1	6/0	5/1
Success rate	100%	78%	83%	100%	88%

3.3. Results for Mixed Sample Type for Training and Testing

A combination of parameters, for example flux+centroid and centroid+roll-off were tested for different voice and noise signals. The samples were prepared by mixing the in constant increasing ratio of speech and noise. The results indicate that the combination of two frequency parameters improves the recognition rate for mixed signals with high background noise—for samples with a voice share of 30%, the recognition rate increased to 78% compared to 55% or 41% for the individual frequency parameters.

4. Conclusions

As we performed VAD based on spectral parameters of signals, it was possible to differentiate noise with speech signals with a proper threshold selection. It was challenging to find a threshold only with an average value for the parameters selected; however, by combining the positive and negative peak average values, a better distinction was achieved. Linear discriminant analysis with flux and centroid parameters provided the best success rate. The system performance could be enhanced by combining the parameters. A larger training test dataset is recommended to verify the results.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ecsa-8-11248/s1.

Author Contributions: R.J. (Ruchi Jha) carried out the recording of sound samples and analysis of the data. R.J. (Reiner Jedermann) selected adequate sound analysis and data classification methods. W.L. contributed to data evaluation and supervised the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all speakers involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Significant Earthquakes. Available online: http://earthquake.usgs.gov/earthquakes/browse/significant.php (accessed on 10 October 2021).
2. Yochum, S.E.; Goertz, L.A.; Jones, P.H. Case study of the Big Bay Dam failure: Accuracy and comparison of breach predictions. *J. Hydraul. Eng.* **2008**, *134*, 1285–1293. [CrossRef]
3. Zhang, D.; Sessa, S.; Kasai, R.; Cosentino, S.; Giacomo, C.; Mochida, Y.; Yamada, H.; Guarnieri, M.; Takanishi, A. Evaluation of a sensor system for detecting humans trapped under rubble: A pilot study. *Sensors* **2018**, *18*, 852. [CrossRef] [PubMed]
4. Aggelopoulos, E.G.; Karabetsos, E.; Constantinou, P.; Uzunoglu, N. Mobile microwave sensor for detection of trapped human beings. *Measurement* **1996**, *18*, 177–183. [CrossRef]
5. *New Technology Can Detect Heartbeats in Rubble*; California Institute of Technology: Pasadena, CA, USA, 17 September 2013; Available online: http://www.jpl.nasa.gov/news/news.php?release=2013-281 (accessed on 1 April 2020).
6. Jha, R.; Lang, W.; Jedermann, R. B4. 5 Sensory options for earthquake victim recovery. In Proceedings of the SMSI 2020-Sensors and Instrumentation, Online Conference, 6 November 2020; pp. 125–126. [CrossRef]
7. Tanyer, S.G.; Özer, H. Voice activity detection in nonstationary noise. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 478–482. [CrossRef]
8. Jeong-Sik, P.; Jung-Seok, Y.; Yong-Ho, S.; Gil-Jin, J. Spectral energy based voice activity detection for real-time voice interface. *J. Theor. Appl. Inf. Technol.* **2017**, *95*, 4304–4312.
9. Alias, F.; Socoro, J.C.; Sevillano, X. A Review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Appl. Sci.* **2016**, *6*, 143. [CrossRef]
10. Mongia, P.K.; Sharma, R.K. Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual. *J. Comput. Netw. Commun.* **2014**, *2014*, 290147. [CrossRef]
11. IRCAM. *A Large Set of Audio Features for Sound Description*; IRCAM Tech. Report; IRCAM: Paris, France, 2003.
12. Da Records. *Geräusche Audio CD, ASIN B00005OCCT*; Da Records: Singapore, 2001; Volumes 1–3.
13. Ghojogh, B.; Crowley, M. Linear and Quadratic Discriminant Analysis: Tutorial. *arXiv* **2019**, arXiv:1906.02590.
14. Hashimoto, K.; Zen, H.; Nankaku, Y.; Lee, A.; Tokuda, K. Bayesian context clustering using cross-validation for speech recognition. *IEICE Trans. Inf. Syst.* **2011**, *94*, 668–678. [CrossRef]