Affective and Pleasurable Design, Vol. 123, 2024, 64–72
https://doi.org/10.54941/ahfe1004677

Development of High-Precision Emotion Estimation Method Using Speech Sound Information With Environmental Noise Reduction and Low Sampling Rate

Kanji Okazaki[1] and Keiichi Watanuki[1,2]

[1]Graduate School of Science and Engineering, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338–8570 Japan
[2]Advanced Institute of Innovative Technology, Saitama University, 255 Shimo-okubo, Sakura-ku, Saitama-shi, Saitama 338–8570 Japan

ABSTRACT

Current research on emotion estimation demonstrates its feasibility at a reduced sampling rate of 6 kHz, thus moving away from traditional methods that depend on higher sampling rates; however, low sampling rates have not been adequately investigated. In addition, noise factors have been limited to electronic sounds rather than environmental. Therefore, this study explores the development of a high-precision emotion estimation method using spoken speech data, focusing on scenarios with environmental noise and low sampling rates. To suppress noise, the proposed method extracts feature quantities for emotion classification using band-pass filters and stacked autoencoders. However, the construction of a high-precision emotion estimation model with these feature quantities required further investigation. Thus, emotion estimation was investigated using a one-dimensional convolutional neural network. The results showed an emotion estimation accuracy of 94.7%, indicating successful noise control. Future work will build on this research to develop emotion estimation methods using spoken speech data that can be employed even in noisy environments.

Keywords: Low sampling rate, Noise reduction, Speech emotion recognition, One-dimensional convolutional neural network

INTRODUCTION

Emotions play a vital role in communication, and according to Albert Mehrabian's research, with meanings typically being conveyed non-verbally, that is, meaning is conveyed 7% through words, 38% through voice, and 55% through visuals. Therefore, accurately recognizing the emotional state of others through non-verbal aspects of the voice can lead to improved communication. Such improvements can not only enhance productivity at work but also enrich everyday life. Previous research on emotion estimation using spoken speech data has dealt with the detection of emotions from speech in noisy environments. However, the noise in these experiments was limited to electronic sounds and did not sufficiently consider real-world noises such as traffic or air conditioner operating sounds. Furthermore, the benefits of

emotion estimation at low sampling rates have not been adequately explored. High noise resilience in emotion estimation, along with data compaction due to low sampling rates, can not only speed up the estimation process, but also promise applicability across various devices and situations. This study aims to achieve accurate emotion estimation by considering realistic noise environments and achieving data compaction.

DATASETS

Specification

In this study, we utilized the Ryerson audio–visual database of emotional speech and song (RAVDESS) dataset for speech data. The RAVDESS dataset includes recordings from 24 actors—12 males and 12 females—portraying eight emotions: neutral, calm, happy, angry, sad, fearful, disgusted, and surprised. The dataset includes 60 utterances from each actor, creating 1440 files. The files are mono audio files with a 16-bit depth and sampling rate of 48 kHz.

DATA PROCESSING AND VISUALIZATION

Low Sampling Rate Estimation

The emotion detection accuracy was assessed for various sampling rates using librosa of PythonâŁ™s library. Features from log-Mel spectrograms were used as input for a logistic regression model to calculate emotion probabilities. The highest probability value was used to determine the predicted emotion.

Speech recognition typically uses 16 kHz data, whereas emotion detection often uses 11–12 kHz. Figure 1 shows a sharp decline in accuracy below 1 kHz. Therefore, in this study, the analysis was conducted at 6 kHz, which is approximately half of the sampling rate commonly used in current practice of emotion estimation.

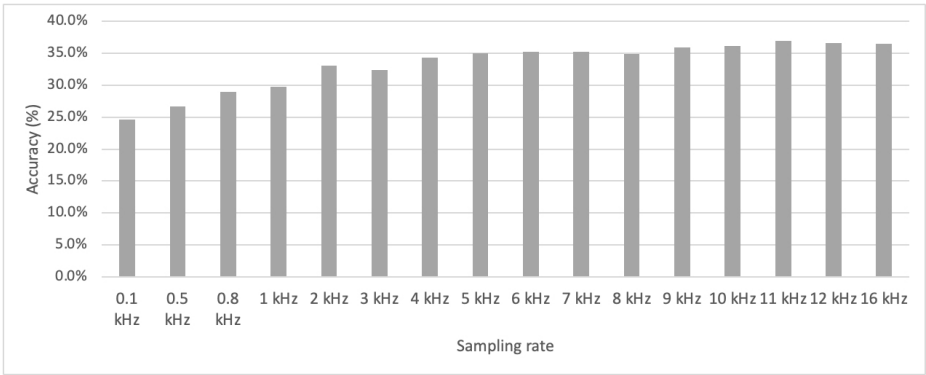

Figure 1: Accuracy variation for emotion estimation due to different sampling rates.

Processing and Visualization

Speech data were randomly selected from the RAVDESS to analyze the features. The speech data features were extracted via fast Fourier transform (FFT). The greatest power is observed to be present in the frequency range dataset of 50–60 Hz, as shown in Figure 2.

As an example of environmental noise, the features of air conditioner operating noise were examined. The results are shown in Figure 3.

The features of speech data overlaid with environmental noise (the operating noise of an air conditioner) are shown in Figure 4.

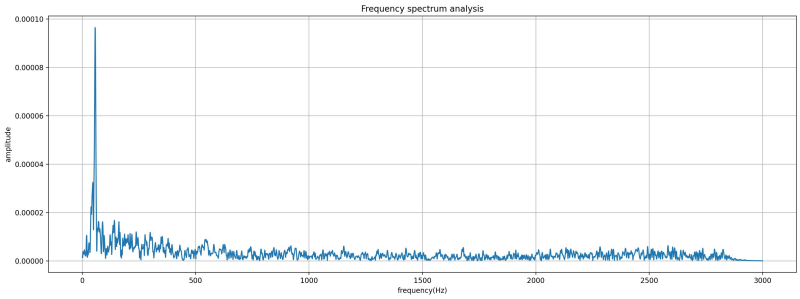

Figure 2: Spectral frequencies of speech data calculated using FFT.

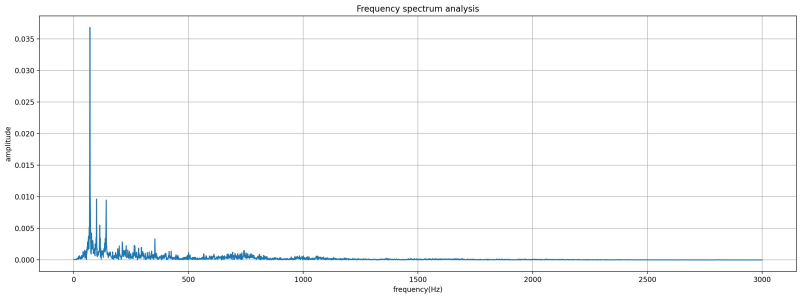

Figure 3: Spectral frequencies of air conditioner operating noise calculated using FFT.

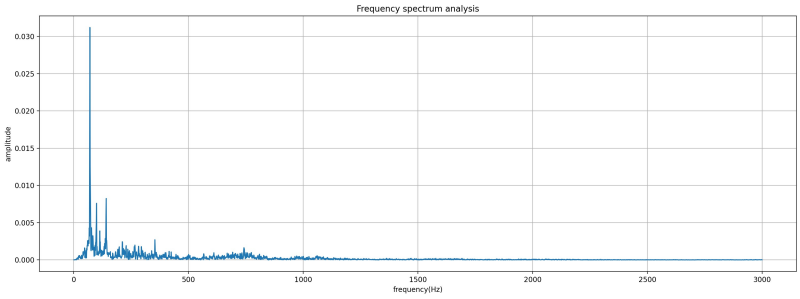

Figure 4: Spectral frequencies of noise-layered speech data calculated using FFT.

Augmentation and Splitting

In this study, we utilized three types of environmental noises: air conditioners, tabletop mini fans, and automobile running noises, which are typical in settings such as hospitals and schools. These environmental noises were combined with the RAVDESS speech data to generate audio samples. The noise volume was normalized and then reduced by 5–15 decibels to match the speech volume (see Figure 5). The noise overlay started randomly within the first 1000 ms. The final set contained 4320 samples with noise (Group A) and 4320 samples without noise (Group B).

As shown in Figure 6, 70% of the data from groups A and B were used for training, with 15% allocated each for validation and testing. The sampling rate was converted to 6 kHz during partitioning. In addition, a bandpass filter (100–2000 Hz) was applied to reduce environmental noise, and melt spectrograms were used to extract speech the features.

MODEL ARCHITECTURE

Stacked Autoencoder

We sought to reduce environmental noise by using an autoencoder to extract emotion-related feature vectors. A stacked autoencoder (see Figure 7) with dataset B was employed for both input and output data.

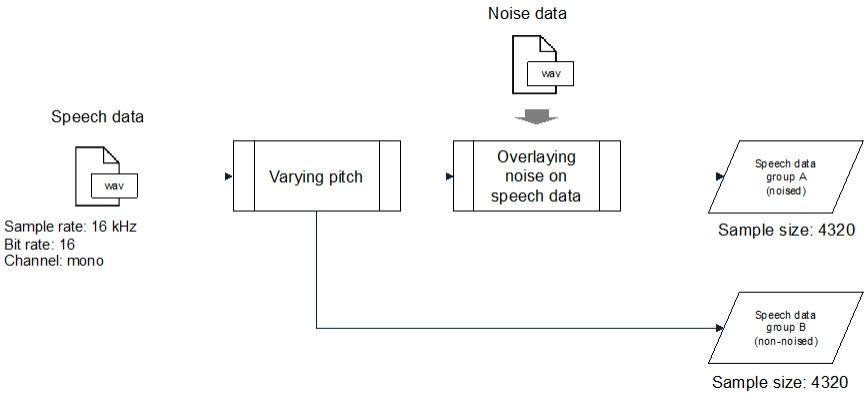

Figure 5: Process of data augmentation.

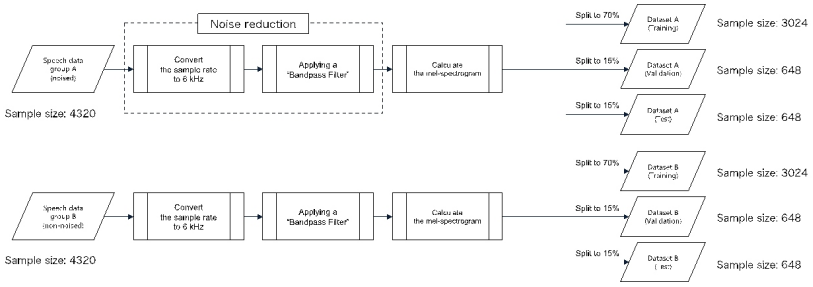

Figure 6: Process of splitting data.

Training

The number of epochs was set to 1000, the batch size was 128, adaptive moment estimation (Adam) was used as the optimization algorithm utilized, and mean squared error (MSE) was used as the loss function.

Results

Dataset A, which included environmental noise, was used in the stacked autoencoder for feature vector estimation. Subsequently, these 16-dimensional vectors were employed to predict emotions using logistic regression. For dataset B, which lacked noise, a matching rate of 36.4% was achieved.

ONE-DIMENSIONAL CONVOLUTIONAL NEURAL NETWORK (1D-CNN)

Data Augmentation and Splitting

The feature vectors obtained using the autoencoders were found to contain a certain level of information for emotion estimation. However, to further improve the accuracy of the emotion estimation, we attempted additional data augmentation and a 1D-CNN (see Figures 8 and 9). Speech data were mixed with mild noise to allow the spoken voice to remain audible.

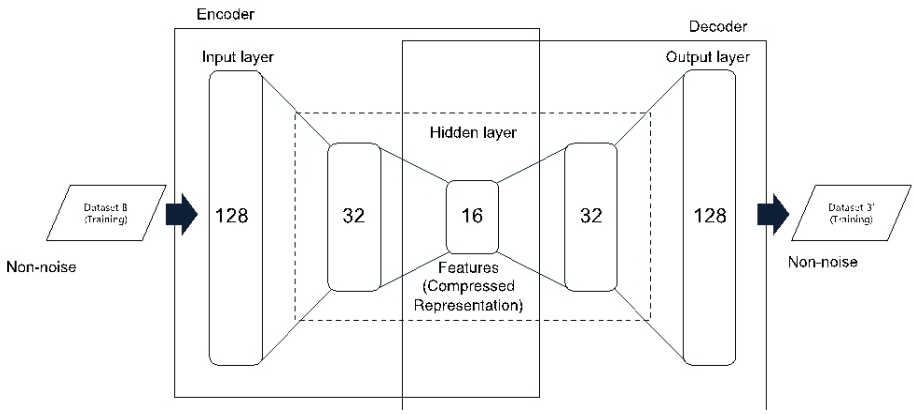

Figure 7: Structure of a stacked autoencoder.

Table 1. Accuracy of the test data.

Input data	Accuracy (%)
Dataset A (test) with noise	**23.9%**
Dataset B (test) without noise	**36.4%**

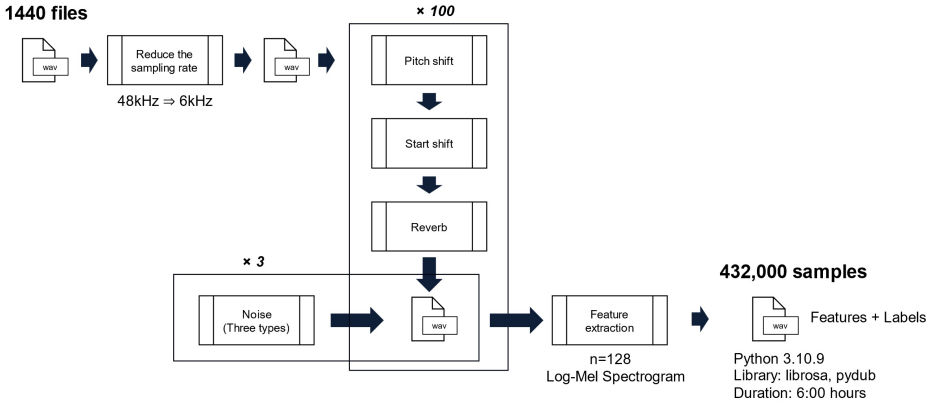

Figure 8: Process of data augmentation.

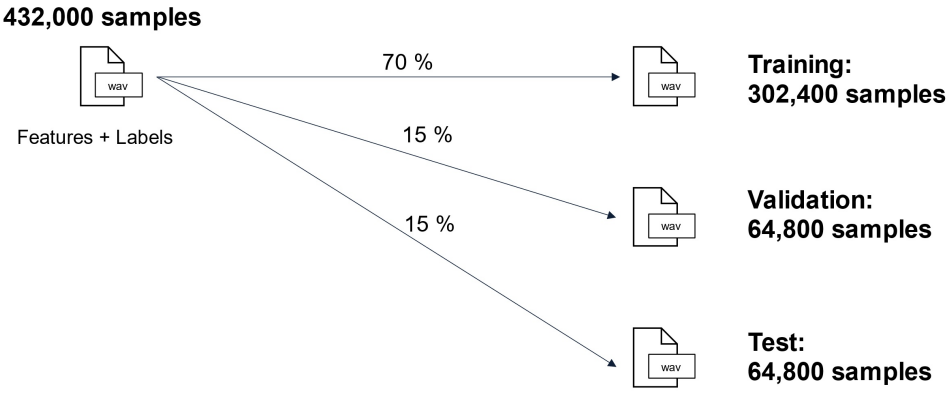

Figure 9: Data splitting for training, validation, and test.

1D-CNN Structure

Using a 1D-CNN, we condensed the necessary information for emotion estimation from speech data with noise. By condensing the information, we aimed to verify whether it had the same effect as noise reduction. Furthermore, we attempted to generalize the accuracy of emotion estimation across various speech data using max pooling and sought to prevent overfitting with dropout. The output utilized a softmax function (see Figure 10). Figure 11 shows the learning process of the 1D-CNN.

Results

The settings of the model parameters are listed in Table 2. The accuracy of the test data is 94.7%.

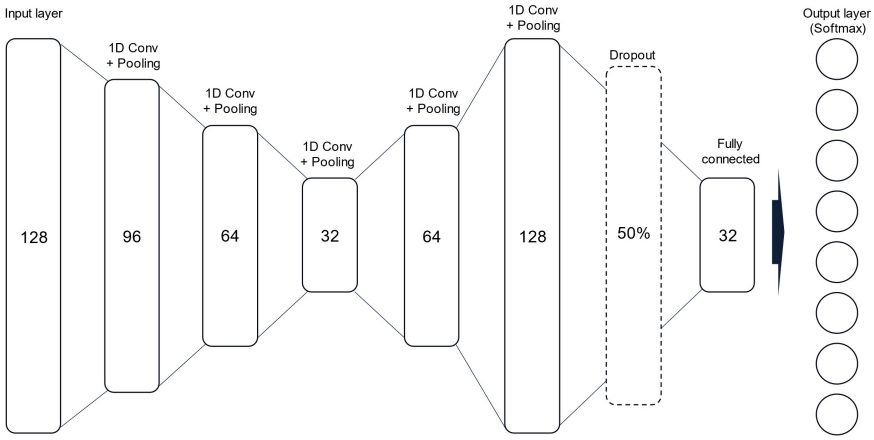

Figure 10: Structure of the 1D-CNN.

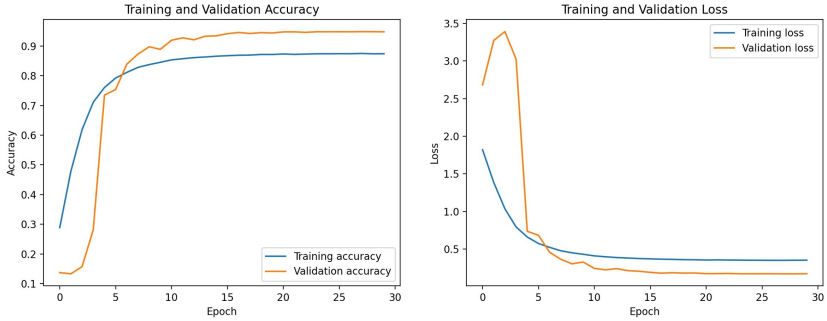

Figure 11: Training and validation accuracy/loss.

Table 2. Model parameters and accuracy.

Parameters	Accuracy
Learning rate: 0.01 Loss function: categorical cross-entropy Optimizer: Adam Epochs: 30 Batch size: 2048	94.7%

CONCLUSION

In this study, we evaluated the possibility of emotion estimation using speech data sampled at a low rate with reduced environmental noise. In practice, it is common to use audio data with a sampling rate of 11 kHz or higher, we demonstrated the feasibility of emotion estimation, even at a low sampling rate of 6 kHz. The study revealed that feature vectors extracted using

stacked autoencoders contained sufficient information for emotion estimation. In addition, dimensionality reduction through autoencoders at a low sampling rate could lead to a reduction in speech data capacity, which is beneficial for achieving high-frequency and high-speed measurements. However, the reduction in environmental noise requires alternative methods to be considered for bandpass filters and stacked autoencoders. In this study, rather than using band-pass filters for noise reduction, a 1D-CNN was adopted for emotion estimation. By condensing the information necessary to estimate emotions with a 1D-CNN, we aimed to reduce the noise contained in the audio data. The results were highly satisfactory with an estimation accuracy of 94.7% for the test data. This research addressed mild noise; however, future research will address the advancement of high-accuracy emotion estimation by targeting noisy environments (such as busy roadsides, train stations, and construction sites). In addition, further research on noise processing using autoencoders is envisioned.

ACKNOWLEDGMENT

We express our deepest gratitude to our colleagues for their generous feedback and profound insight into the formulation of this study. Their contributions have significantly enhanced the quality of our research and guided us toward a more refined and insightful study. We extend our sincere gratitude and appreciation for their invaluable support.

REFERENCES

Badshah, A. M., Ahmad, J., Rahim, N., Baik, S. W. (2017) "Speech emotion recognition from spectrograms with deep convolutional neural network", in Proc. Int. Conf. Plat-form Technol. Service (PlatCon), Busan, Korea (South), pp. 1–5.

Boashash, B., ed. (2003) Time-Frequency Signal Analysis and Processing: A Comprehensive Reference. Oxford: Elsevier Science.

Bracewell, R. N. (2000) The Fourier Transform and its Applications. Boston: McGraw-Hill.

Dossou, B. F. P., Gbenou, Y. K. S. (2021) "FSER: Deep Convolutional Neural Networks for Speech Emotion Recognition", 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 3526–3531.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., Herrera, F. (2011) "An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes", Pattern Recognition, Vol. 44, No. 8, pp. 1761–1776.

Issa, D., Demirci, M. F., Yazici, A. (2020) "Speech emotion recognition with deep convolutional neural networks", Biomedical Signal Processing and Control, Vol. 59, p. 101894.

Livingstone, S. R., Russo, F. A. (2018) "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)", Funding Information Natural Sciences and Engineering Research Council of Canada: 2012–341583 Hear the world research chair in music and emotional speech from Phonak.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O. (2015) "librosa: Audio and music signal analysis in python", in Proceedings of the 14th python in science conference, Vol. 8, pp. 18–25.

Mehrabian, A. (1971) Silent Messages. Belmont, California: Wadsworth Publishing Company.
Rifkin, R., Klautau, A. (2004) "In defense of one-vs-all classification", The Journal of Machine Learning Research, Vol. 5, pp. 101–141.
Sadok, S., Leglaive, S., Seguier, R. (2023) "A vector quantized masked autoencoder for speech emotion recognition", arXiv:2304.11117 [cs, eess] version: 1. [Online]. Available: https://arxiv.org/abs/2304.11117
Venkataramanan, K., Rajamohan, H. R. (2019) Emotion recognition from speech.