Affective and Pleasurable Design, Vol. 164, 2025, 32–41
https://doi.org/10.54941/ahfe1006058

Development of a Fast and High-Precision Audio Noise Reduction System to Enhance the Accuracy of Emotion Estimation in Practical Applications

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

[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

Speech-based emotion estimation has diverse applications, including mental health monitoring, human–computer interaction, and communication enhancement. The accurate estimation of emotions from speech is crucial in the detection of psychological stress, which is a growing concern in today's high-stress societies. However, environmental noise significantly degrades estimation accuracy, and studies focusing on noise reduction specifically optimized for emotion estimation remain scarce. This study evaluated the impact of noise reduction on emotion estimation by comparing traditional signal processing (spectral subtraction, Wiener filtering) with deep learning-based methods (U-Net autoencoder, convolutional autoencoder). The effectiveness of each method is examined under continuous vehicle driving and transient clapping noise. The results indicate that traditional techniques effectively suppress continuous noise but struggle with transient noise, whereas AE-based methods, particularly U-Net autoencoder, significantly enhance the estimation accuracy in complex noise environments. This study underscores the importance of emotion-aware noise reduction and suggests that deep learning-based denoising techniques can significantly improve real-world applications. Future research will focus on further optimizing the AE architectures and integrating them into real-time systems.

Keywords: Speech emotion estimation, Noise reduction, U-Net, Spectral subtraction

INTRODUCTION

Understanding speakers' emotions is crucial for smooth communication, interpersonal relationships, and psychological stress management. Advances in emotion estimation technology have enabled speech-based inferences of psychological states, highlighting their potential for stress assessment and mental health support. As psychological stress affects many individuals, effective stress management remains a major societal challenge. Technologies

that estimate stress levels and enable early intervention can benefit the healthcare, welfare, education, and workplace environments.

Various emotion estimation methods exist, including facial expression estimation, EEG analysis, and heart rate variability (HRV) measurements. However, speech-based methods offer a distinct advantage, as they require no specialized equipment and can be performed by anyone. Speech-based emotion estimation using smartphones and microphone-equipped devices is one of the most accessible real-world technologies. However, environmental noise significantly degrades estimation accuracy, making effective noise suppression essential. Conventional noise reduction techniques, such as spectral subtraction and Wiener filtering, have been extensively studied. However, these techniques primarily focus on speech enhancement rather than optimization for emotion estimation.

To address this, we compared conventional noise reduction methods (e.g., spectral subtraction and Wiener filtering) with a deep-learning-based approach using autoencoders (AEs) and assessed their impacts on emotion estimation accuracy. Each method was evaluated under continuous vehicle and transient clapping noise conditions. The results confirm that AE-based noise reduction enhances the emotion estimation accuracy compared with conventional methods. In particular, AEs exhibit high adaptability to nonstationary noise, which ensures stable estimation, even in noisy environments. This study advances speech-based emotion estimation and lays the foundation for its practical applications in stress management and mental healthcare.

RELATED WORK

This study enhances the emotion estimation accuracy in low-sampling rate and noisy environments by comparing conventional and deep learning-based noise reduction methods. This section reviews conventional noise-reduction techniques, deep-learning-based approaches, and emotion estimation in low-sampling-rate environments.

Conventional Noise Reduction Techniques

Noise reduction techniques for speech signals have been widely studied. For example, spectral subtraction and Wiener filtering are commonly used. Spectral subtraction removes an estimated noise spectrum via a short-time Fourier transform (STFT) and is effective against stationary noise, such as engine sounds. However, excessive subtraction may introduce musical noise and thereby distort speech. Wiener filtering adaptively reduces noise based on the signal-to-noise ratio (SNR) and thereby preserves speech clarity better than does spectral subtraction. However, SNR estimation errors may cause unintended speech removal. Although these methods are computationally efficient and suitable for real-time processing, they struggle with nonstationary noise, such as clapping sounds and human speech.

Deep Learning-Based Noise Reduction

In recent years, deep learning-based noise reduction has gained attention, with autoencoder (AE)-based methods demonstrating high performance. The U-Net autoencoder (U-Net AE) extends the U-Net architecture from image processing to speech signal processing by using skip connections to preserve high-frequency components while removing noise, making it particularly effective for nonstationary noise, such as clapping and keyboard typing sounds. The convolutional autoencoder (CAE) has a simpler encoder–decoder structure without skip connections that enables the efficient learning of complex noise patterns while reducing the computational cost, although it may lose high-frequency information. Moreover, CAE with data augmentation further improves adaptability by applying pitch shifting, time stretching, and reverberation effects, thereby enhancing noise robustness beyond conventional CAE methods.

Emotion Estimation in Low-Sampling Rate Environments

Most emotion estimation studies assume high-sampling-rate environments (e.g., 11.25 kHz, 16 kHz), whereas research conducted under low-sampling-rate settings (e.g., 6 kHz) remains limited. Recent studies have suggested that emotion estimation at 6 kHz is feasible, and an accuracy of 94.7% was achieved via low-pass filters and noise-included data augmentation that suppressed environmental noise while preserving features. This study further demonstrates that a 1D-CNN outperforms conventional methods in terms of emotion estimation accuracy.

Research Objective

This study compared conventional noise reduction methods (spectral subtraction, Wiener filtering) with deep learning-based techniques (U-Net AE and CAE) to evaluate their impacts on speech emotion estimation accuracy. This study assesses the effectiveness of each method under environmental noise and low-sampling-rate conditions to determine its practicality. The goal of this study is to identify the strengths and weaknesses of both approaches and propose an optimal noise-reduction strategy based on continuous and transient noise types.

PROPOSED METHOD

This study aims to improve the speech emotion estimation accuracy in noisy environments by comparing and evaluating multiple noise reduction techniques. The evaluated methods can be broadly categorized into conventional signal-processing techniques and deep-learning-based approaches.

Conventional Noise Reduction Methods:

- Spectral subtraction
- Wiener filtering

Deep Learning-Based Noise Reduction Methods:

- U-Net autoencoder (U-Net AE)
- Convolutional autoencoder (CAE)

Spectral Subtraction

Spectral Subtraction is a noise-reduction technique that estimates and subtracts noise components from the spectral representations of speech signals. The noise-removal process followed the following steps:

1. A short-time Fourier transform (STFT) was applied to obtain the magnitude and phase spectra of the speech signal.
2. The noise spectrum was estimated by averaging the first 10 frames, assuming that they mostly contained noise.
3. The estimated noise spectrum was subtracted from the magnitude spectrum (negative values were clipped to zero).
4. An inverse STFT (ISTFT) was applied while preserving the original phase spectrum to reconstruct the time-domain signal.

This method is computationally efficient and effective for continuous noise. However, it struggles with transient noise. Additionally, excessive noise subtraction may introduce musical noise and thereby lead to speech distortion.

Wiener Filtering

Wiener filtering is a statistical noise reduction technique that applies an optimal filter based on the signal-to-noise ratio (SNR). The noise-removal process followed the following steps.

1. An STFT was applied to obtain the magnitude and phase spectra of the speech signal.
2. The noise power spectrum was estimated by averaging the power spectrum over the first 10 frames, assuming that they primarily contained noise.
3. Compute the Wiener gain for each frequency band to suppress noise components adaptively.
4. Apply an ISTFT to reconstruct the time-domain speech signal.

Compared with spectral subtraction, this method provides adaptive noise suppression. However, it tends to remove high-frequency components of speech, which may lead to degraded speech quality owing to excessive noise reduction.

U-Net Autoencoder

To maintain the emotion estimation performance, even in noisy environments, this study introduces an autoencoder (AE) with a U-Net architecture. U-Net is a network that extends the encoder–decoder structure by incorporating skip connections and consists of the following components.

Encoder:

- Convolutional layers (Conv1D): Extracts features using filters with sizes of 16, 32, 64, and 128.
- Downsampling: Uses stride-2 convolution to compress dimensionality and extract key features.
- Bottleneck layer: 256-filter convolutional layer that captures the most abstract feature representations.

Decoder:

- Transposed convolutional layers (Conv1DTranspose): Restore the temporal resolution of the original speech signal using filters with sizes 128, 64, 32, and 16.
- Skip connections merge feature maps from the encoder to prevent the loss of high-frequency information.
- Output layer: Single-channel convolutional layer with linear activation that generates a denoised speech signal.

This model is adaptable to complex noisy environments in which spectral subtraction and Wiener filtering are challenging. Moreover, the proposed model is highly effective against transient noises, such as clapping sounds.

Convolutional Autoencoder

As an alternative to U-Net, this study evaluated a convolutional autoencoder (CAE) that does not include skip connections. The CAE has the following characteristics.

Encoder:

Three Conv1D layers with 16, 32, and 64 filters for feature extraction.

Decoder:

Three Conv1DTranspose layers with 64, 32, and 16 filters were used to reconstruct temporal resolution.

Output Layer:

A single-channel Conv1D layer with linear activation is used to generate the denoised speech signal.

Compared with U-Net, CAE has a simpler structure and lower computational cost. However, because it lacks skip connections, it may have a reduced capability to restore high-frequency information.

Key Features and Expected Impact of the Proposed Method

- Conventional methods, such as spectral subtraction and Wiener filtering, are effective for detecting stationary noise (e.g., vehicle driving noise).
- The U-Net Autoencoder demonstrated the highest effectiveness for transient noise (e.g., clapping sounds).
- Combining CAE with data augmentation has the potential to enhance adaptability to a wider range of noise environments.

This study compares these methods and examines the optimal noise reduction approach that contributes to improving speech emotion estimation accuracy.

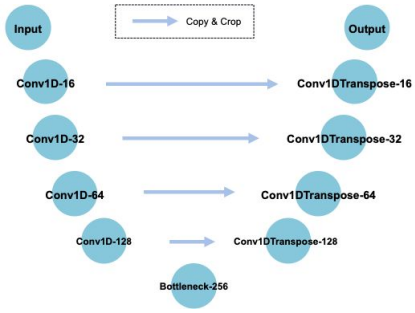

Figure 1: Structure of U-Net autoencoder (U-Net AE).

EXPERIMENTS AND EVALUATION

Experimental Setup

To evaluate the impacts of different noise reduction methods on speech emotion estimation, experiments were conducted in two noisy environments.

- Vehicle driving noise: Environmental noise recorded inside a moving vehicle.
- Clapping noise: Hand clapping sounds recorded in an arena.

The evaluated noise reduction methods are as follows.
Conventional Signal Processing Methods:

- Spectral subtraction
- Wiener filtering

Deep Learning-based Methods:

- U-Net autoencoder (U-Net AE)
- Convolutional autoencoder (CAE)

Performance Metrics

The following metrics were used to assess both the quality of noise reduction and its impact on emotion estimation accuracy:
Noise Reduction Performance:

- Signal-to-noise ratio (SNR)
- Signal-to-distortion ratio (SDR)
- Short-time objective intelligibility (STOI)
- Perceptual evaluation of speech quality (PESQ)

Emotion Estimation Performance:

- Emotion classification accuracy using a 1D-CNN model trained on the RAVDESS dataset.
- Pre- and post-denoising speech samples were compared to analyze the effect of each noise reduction method on the emotion estimation accuracy.

RESULTS AND DISCUSSION

This study compared the performance of multiple noise reduction methods in two distinct noise environments: vehicle-driving noise and clapping noise. The evaluation was conducted using the following metrics: SDR, STOI, and PESQ. In addition, emotion estimation accuracy was analyzed using speech data after noise reduction.

Comparison of Noise Reduction Performance

Vehicle Driving Noise Environment

Spectral subtraction achieved the highest SDR (10.15 dB) and STOI (0.93) values, demonstrating its effectiveness in reducing vehicle driving noise. The U-Net Autoencoder also exhibited relatively good performance, achieving a high PESQ score (1.54). The CNN autoencoder exhibited inferior performance compared with the other methods.

Clapping Noise Environment

The U-Net Autoencoder achieved the highest SDR (8.41 dB) value, confirming its effectiveness in removing the clapping noise. Spectral subtraction and Wiener filtering exhibited high STOI and PESQ values; however, their low SDR values indicated that they were less effective at reducing clapping noise. The CNN Autoencoder showed some effectiveness but did not perform as well as the U-Net Autoencoder.

Comparison of Emotion Estimation Performance

Vehicle Driving Noise Environment

Spectral subtraction maintained the most stable emotion distribution, with particularly high proportions of calm, neutral, and sad emotions. The CNN autoencoder classified happy at a high rate of 56.04%, whereas the CAE with data augmentation excessively emphasized surprise (84.74%). Compared with clean speech, all methods showed a significant decrease in the proportion of happy speech.

Clapping Noise Environment

Spectral subtraction and Wiener filtering tended to overemphasize calmness, reaching 74.00%. The CNN autoencoder significantly amplified the anger, disgust, and surprise signals. Fearfulness was highest in the CNN autoencoder, suggesting that a clapping noise may influence the perception of fear.

Table 1: Noise reduction performance in the vehicle driving noise environment.

Method	SNR (dB)	SDR (dB)	STOI	PESQ
Spectral Subtraction	10.18	10.15	0.93	1.83
Wiener Filtering	8.26	8.18	0.92	1.64
U-Net AE	8.36	9.14	0.87	1.54
CNN AE	6.63	6.83	0.85	1.35

Table 2: Noise reduction performance in the clapping noise environment.

Method	SNR (dB)	SDR (dB)	STOI	PESQ
Spectral Subtraction	4.33	4.10	0.89	1.33
Wiener Filtering	3.61	3.66	0.89	1.31
U-Net AE	7.81	8.41	0.85	1.23
CNN AE	6.28	6.26	0.83	1.22

Table 3: Emotion estimation performance in the vehicle driving noise environment.

File Name	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise
Clean Audio	2.97	0.04	1.51	0.62	93.89	0.10	0.20	0.67
Noisy Audio	8.39	0.79	28.43	24.03	10.92	1.14	6.27	20.03
Spectral Subtraction	0.95	34.85	3.62	7.41	5.12	15.33	18.80	13.91
Wiener Filtering	0.33	29.55	2.74	25.03	5.46	14.34	16.17	6.38
U-Net AE	0.00	0.02	0.00	0.11	0.03	0.02	99.63	0.18
CNN AE	4.33	0.09	6.73	3.96	56.04	0.32	21.19	7.33

Table 4: Emotion estimation performance in clapping noise environment.

File Name	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprise
Clean Audio	2.97	0.04	1.51	0.62	93.89	0.10	0.20	0.67
Noisy Audio	75.00	12.69	2.49	3.08	2.18	3.19	0.24	1.12
Spectral Subtraction	9.47	74.00	2.89	1.15	2.34	8.90	0.32	0.92
Wiener Filtering	19.77	59.40	5.36	1.57	1.44	10.61	0.78	1.07
U-Net AE	2.46	1.41	6.40	5.93	6.46	1.37	58.43	17.54
CNN AE	28.84	1.13	29.67	6.96	2.13	0.43	0.62	30.21

CONCLUSION AND FUTURE WORK

Conclusion

This study examined the impact of noise reduction methods on speech emotion estimation in vehicular driving and clapping noise environments. They compared conventional methods (spectral subtraction, Wiener filtering) with deep learning-based approaches (U-Net AE and CAE) using SDR, STOI, and PESQ.

Key findings include:

- Spectral subtraction was the most effective method for vehicle noise and achieved the highest SDR (10.15 dB) and STOI (0.93) values.

- The U-Net Autoencoder performed the best for clapping noise, with the highest SDR (8.41 dB).
- Wiener filtering provided moderate noise reduction but introduced speech distortion.
- CAE-based methods effectively handled complex noise but impacted speech clarity.

In emotion estimation:

- Spectral subtraction preserved the most natural emotion distribution in vehicle noise environments.
- CNN autoencoder amplified angry and disgust emotions in clapping noise environments.
- All methods reduced the proportion of happy emotions compared to clean speech.

These results highlight the significant impact of noise reduction choices on both speech enhancement and emotion estimation accuracy, underscoring the need to select methods based on the noise type.

Future Work

Future research will focus on:

- Improving deep learning-based noise reduction models for better generalization across noise environments.
- Developing real-time implementations for practical use in speech-based emotion estimation.
- Exploring adaptive noise reduction techniques that dynamically adjust to varying noise conditions.
- Expanding evaluations with diverse datasets to validate effectiveness in real-world scenarios.

These efforts aim to advance noise-robust speech emotion estimation and support applications in mental health monitoring, human–computer interaction, and communication support systems.

ACKNOWLEDGMENT

We sincerely thank our colleagues for their invaluable feedback and support, which significantly enhanced this study. We also extend our gratitude to Haruka Kudo, a representative of SHIN4NY Inc., for her valuable advice and assistance in collecting environmental noise data.

REFERENCES

Benesty, J., Chen, J., Huang, Y. and Doclo, S. (2005) 'Study of the Wiener filter for noise reduction', in *Speech Enhancement, Signals and Communication Technology*. Berlin, Heidelberg: Springer. doi: 10.1007/3-540-27489-8_2.

Boll, S. (1979a) 'A spectral subtraction algorithm for suppression of acoustic noise in speech', *ICASSP '79 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Washington, DC, USA, pp. 200–203. doi: 10.1109/ICASSP.1979.1170696.

Boll, S. (1979b) 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), pp. 113–120. doi: 10.1109/TASSP.1979.1163209.

Chen, J., Benesty, J., Huang, Y. and Doclo, S. (2006) 'New insights into the noise reduction Wiener filter', *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), pp. 1218–1234. doi: 10.1109/TSA.2005.860851.

El Ayadi, M., Kamel, M. S. and Karray, F. (2011) 'Survey on speech emotion estimation: Features, classification schemes, and databases', *Pattern Recognition*, 44(3), pp. 572–587. doi: 10.1016/j.patcog.2010.09.020.

Gobl, C. and Chasaide, A. N. (2003) 'The role of voice quality in communicating emotion, mood and attitude', *Speech Communication*, 40(1–2), pp. 189–212. doi: 10.1016/S0167-6393(02)00082-1.

Kamath, S. and Loizou, P. (2002) 'A multi-band spectral subtraction method for enhancing speech corrupted by colored noise', *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, pp. IV-4164-IV-4164. doi: 10.1109/ICASSP.2002.5745591.

Okazaki, K. and Watanuki, K. (2024b) 'Development of high-precision emotion estimation method using speech sound information with environmental noise reduction and low sampling rate', *AHFE 2024*.

Riahi, A. and Plourde, É. (2023) 'Single channel speech enhancement using U-Net spiking neural networks', *IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, Regina, SK, Canada, pp. 111–116. doi: 10.1109/CCECE58730.2023.10288830.

Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional networks for biomedical image segmentation', *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference*, Munich, Germany, October 5-9, 2015. Springer International Publishing.

Singh, L. and Sridharan, S. (1998) 'Speech enhancement using critical band spectral subtraction', *ICSPLP*.

Stahl, V., Fischer, A. and Bippus, R. (2000) 'Quantile based noise estimation for spectral subtraction and Wiener filtering', *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, pp. 1875–1878. doi: 10.1109/ICASSP.2000.862122.

Tamura, S. and Waibel, A. (1988) 'Noise reduction using connectionist models', *ICASSP-88 International Conference on Acoustics, Speech, and Signal Processing*, New York, NY, USA, pp. 553–556. doi: 10.1109/ICASSP.1988.196643.

Wiener, N. (1949) *Extrapolation, interpolation, and smoothing of stationary time series: With engineering applications*. Cambridge: MIT Press.

Xia, B. and Bao, C. (2014) 'Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification', *Speech Communication*, 60, pp. 13–29.