Investigation of noise-reverberation-robustness of modulation spectral features for speech-emotion recognition

Taiyang Guo, Sixia Li, Masashi Unoki and Shogo Okada
Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923–1292 Japan
E-mail: guotaiyang@jaist.ac.jp, okada-s@jaist.ac.jp

***Abstract*—Speech-emotion recognition (SER) in noisy reverberant environments is a fundamental technique for real-world applications, including call center service and psychological disease diagnosis. However, in daily auditory environments with noise and reverberation, previous studies using acoustic features could not achieve the same emotion-recognition rates as in an ideal experimental environment (with no noise and no reverberation). To remedy this imperfection, it is necessary to find robust features against noise and reverberation for SER. However, it has been proved that a daily noisy reverberant environment (signal-to-noise ratio is greater than 10 dB and reverberation time is less than 1.0 s) does not affect humans' vocal-emotion recognition. On the basis of the auditory system of human perception, previous research proposed modulation spectral features (MSFs) that contribute to vocal-emotion recognition by humans. Using MSFs has the potential to improve SER in noisy reverberant environments. We investigated the effectiveness and robustness of MSFs for SER in noisy reverberant environments. We used noise-vocoded speech, which is synthesized speech that retains emotional components of speech signals in noisy reverberant environments as speech data. We also used a support vector machine as the classifier to carry out emotion recognition. The experimental results indicate that compared with two widely used feature sets, using MSFs improved the recognition accuracy in 13 of the 26 environments with an average improvement of 11.38%. Thus, MSFs contribute to SER and are robust against noise and reverberation.**

I. Introduction

In speech-processing research, machine-learning techniques have enabled computers to solve many speech-recognition tasks as effectively as humans, such as the great success achieved in automatic speech recognition [1]–[3]. To create a user-friendly human-machine interface, it is not enough to understand what the user said but also the user's emotions [4], [5]. Accurately recognizing emotional information is an indispensable part of smooth communication. Accordingly, speech-emotion recognition (SER) has become an important task in speech processing.

Extracting suitable features to precisely describe the emotion information is a core part of SER. By using the widely used Mel frequency cepstral coefficient (MFCC) [6], [7], several studies achieved high emotion-recognition rates in ideal experimental environments without disturbances such as noise. Researchers have been paying more attention to the problem of speech recognition in real auditory environments. In such environments, noise and reverberation always exist; such disturbances often affect the perception of speech. As shown in previous research, in an environment containing background noise and room reverberation [8], the emotion-recognition rates significantly decrease compared with clean test data. Various solutions have been proposed to mitigate such limitations from noise and reverberation. However, even modified speech-recognition systems or enhanced algorithms [9], [10] could not achieve the same emotion-recognition rates as in ideal auditory environments.

To solve the lack of robustness against noise and reverberation of SER, one study on vocal-emotion recognition using noise-vocoded speech (NVS) [11] provided the clue to extract robust features against noise and reverberation from the perspective of human speech perception. The experimental results indicated that vocal-emotion recognition is not affected by daily noise and reverberation conditions (signal-to-noise ratio (SNR) is greater than 10 dB and reverberation time is less than 1.0 s). Thus, it is necessary to extract important features in vocal-emotion recognition. Based on evidence obtained from both modern physiological [12] and psychological [13] models, an auditory filterbank exists in the human auditory system. The auditory filterbank decomposes speech signals into channel signals (temporal amplitude envelope (TAE) and temporal fine structure) in the time-frequency domain. Several studies have proved that the TAE and its modulation cues play an important role in speech recognition [14–16]. One study on vocal-emotion recognition [17] proposed modulation spectral features (MSFs) on the basis of the modulation analysis of the TAE and proved that MSFs contribute to vocal-emotion recognition. Due to vocal-emotion recognition being robust against noise and reverberation, as important features for vocal-emotion recognition, MSFs have the potential to improve the emotion-recognition rate in noisy reverberant environments in SER. However, how MSFs perform and whether they have good noise-reverberation robustness is still unclear.

We focused on the effectiveness and robustness of MSFs for SER in noisy reverberant environments. We used a support vector machine (SVM) as the classifier to carry out emotion

978-616-590-477-3 ©2022 APSIPA APSIPA ASC 2022

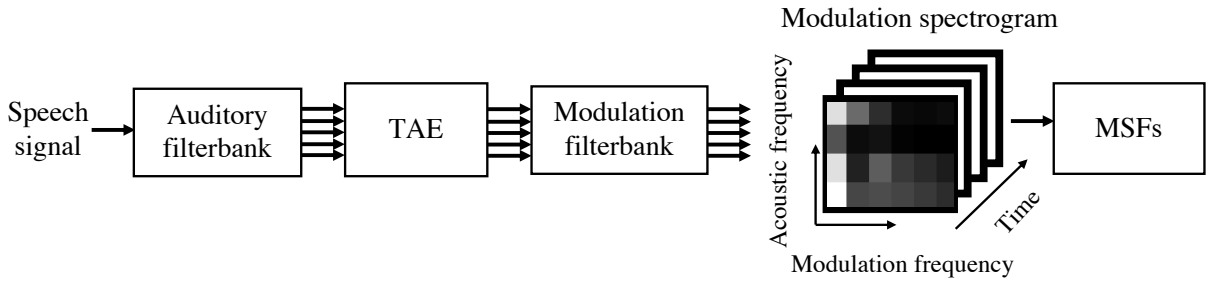


Fig. 1. Process of extracting MSFs [17].

classification to investigate the following two research questions: (1) Whether MSFs contribute to SER and (2) Whether MSFs have good noise-reverberation-robustness for SER.

II. IMODULATION SPECTRAL FEATURE EXTRACTION

We used the same MSF-extraction method as in a previous study [17]. To extract MSFs from emotional speech signals, it is first necessary to calculate the modulation spectrogram using a modulation filterbank. Figure 1 shows the modulation process we used. Emotional speech signals s were divided into several frequency bands by using an auditory-based band-pass filterbank:

$$s_k(n) = s(n) * h_k(n), \tag{1}$$

where * denotes the convolution operator, $h_k(n)$ is the impulse response of the k^{th} channel, and n is the sample number in the time domain. The 6^{th}-order Butterworth infinite impulse response (IIR) band-pass filterbank was used as the auditory filterbank. The bandwidth of each filter was the bandwidth of the human auditory filter, and the order of the filters was determined in accordance with the equivalent rectangular bandwidth ($\mathrm{ERB_N}$) and $\mathrm{ERB_N}$-number scale [18], where the unit of $\mathrm{ERB_N}$-number is Cam. The boundary frequencies of the band-pass filters (BPFs) were defined as $\mathrm{ERB_N}$-number from 3 to 35 Cam with an 8 $\mathrm{ERB_N}$ bandwidth, and the number of channels was 4.

The temporal envelope of the output signal from each BPF $s_k(n)$ was extracted using the Hilbert transformation, and a 2^{nd}-order Butterworth IIR low-pass filter (LPF) (cut-off frequency is 64 Hz) was used as follows:

$$e_k(n) = \mathrm{LPF}\left[|s_k(n) + j\mathcal{H}[s_k(n)]|\right], \tag{2}$$

where $\mathcal{H}$ denotes the Hilbert transform.

The next step involved decomposing the temporal envelope into several modulation-frequency bands by using a modulation filterbank:

$$E_{k,m}(n) = g_m(n) * (e_k(n) - \overline{e_k(n)}), \tag{3}$$

where m is the channel number of the modulation filter, $g_m(n)$ is the impulse response of the modulation filterbank, and $\overline{e_k(n)}$ is the time-averaged amplitude of $e_k(n)$. The modulation filterbank consisted of six filters (one LPF and five BPFs). The boundary frequencies of the filters were spaced on an octave frequency band from 2 to 64 Hz.

The root-mean-square of $E_{k,m}(n)$ is calculated as the modulation spectrogram,

$$\bar{E}_{k,m}(n) = \sqrt{\frac{1}{N}\sum_{n=1}^{N} E_{k,m}^2(n)}, \tag{4}$$

where the N is the length of the speech signal $s(n)$.

In the next step, ten MSFs should be extracted from the modulation spectrograms. They are the MSFs in the acoustic-frequency domain (the subscript is m) and in the modulation-frequency domain (the subscript is k): the modulation spectral centroid ($\mathrm{MSCR}_{m/k}$), modulation spectral spread ($\mathrm{MSSP}_{m/k}$), modulation spectral skewness ($\mathrm{MSSK}_{m/k}$), and modulation spectral kurtosis ($\mathrm{MSKT}_{m/k}$), which are defined as follows:

$$\mathrm{MSCR_m} = \frac{\sum_{k=1}^{K} k\bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}}, \tag{5}$$

$$\mathrm{MSSP_m} = \frac{\sum_{k=1}^{K} [k - \mathrm{MSCR}_m]^2 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}}, \tag{6}$$

$$\mathrm{MSSK_m} = \frac{\sum_{k=1}^{K} [k - \mathrm{MSCR}_m]^3 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}}, \tag{7}$$

$$\mathrm{MSKT_m} = \frac{\sum_{k=1}^{K} [k - \mathrm{MSCR}_m]^4 \bar{E}_{k,m}}{\sum_{k=1}^{K} \bar{E}_{k,m}}, \tag{8}$$

$$\mathrm{MSCR_k} = \frac{\sum_{m=1}^{M} m\bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}}, \tag{9}$$

$$\mathrm{MSSP_k} = \frac{\sum_{m=1}^{M} [m - \mathrm{MSCR}_k]^2 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}}, \tag{10}$$

$$\mathrm{MSSK_k} = \frac{\sum_{m=1}^{M} [m - \mathrm{MSCR}_k]^3 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}}, \tag{11}$$

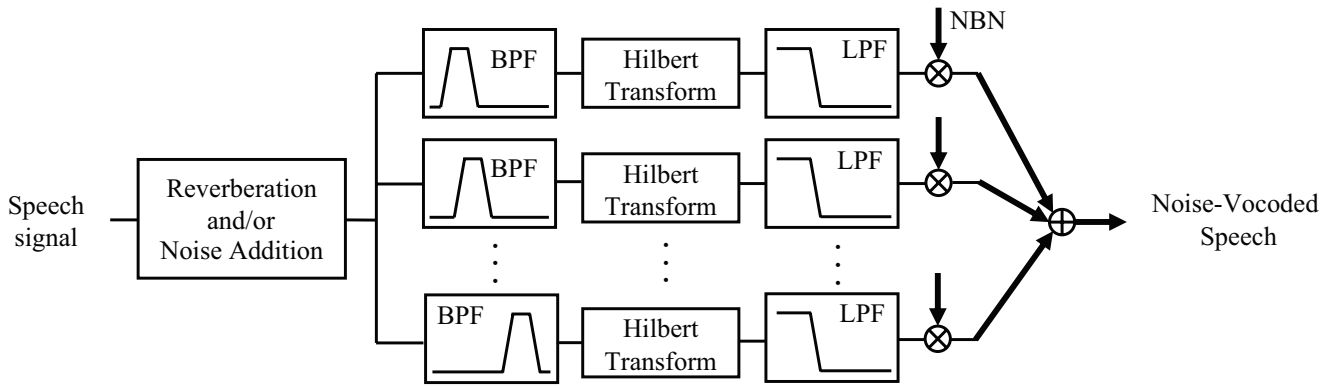


Fig. 2. Schematic diagram of noise-vocoded method used to generate experiment stimuli (NBN: narrow band noise). BPFs were defined as $\mathrm{ERB_N}$-number from 3 to 35 Cam with bandwidths of 2 $\mathrm{ERB_N}$. Number of channels was 16 [11].

$$\mathrm{MSKT_k} = \frac{\sum_{m=1}^{M}[m - \mathrm{MSCR}_k]^4 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}}. \qquad (12)$$

The last two MSFs in the acoustic-frequency and modulation-frequency domains were modulation spectral tilts (MSTL_m and MSTL_k), which are the linear regression coefficients obtained by fitting the first-degree polynomial to the modulation spectrograms.

III. Experimental setting

A. Original speech data

We used the Fujitsu Japanese Emotional Speech Database used by Zhu et al. [11,16] as the original speech data. In this database, a professional actress's sentences are expressed, and each sentence contains one of five emotions (neutral, joy, cold anger, sadness, and hot anger). The speech data were recorded with a sampling frequency of 20 kHz (the original signal was 22.05 kHz but was resampled to 20 kHz to match the conditions of other experiments) and 16-bit quantization, and the duration of each utterance was about 3 to 4 s.

B. Noise-vocoded speech generation

We used the same method as in a previous study [11] to generate NVS in noisy and reverberant environments as speech data. NVS is synthesized speech that preserves the TAE information, which contains important emotion information [11,17]. The experimental stimuli of NVS were synthesized in three environments: noisy, reverberant, and noisy reverberant. The experimental stimuli were created by the following procedure.

To produce noisy emotional speech, we used stationary noise (white Gaussian noise), and the adjusted noise was added to the speech so that the SNR of the original speech and noise would differ. SNRs of 20, 15, 10, 5, 0, and -5 dB were selected.

Concerning reverberant emotional speech, previous research used a statistical room impulse response (Schroeder model) [19]. Five types of room-impulse responses with reverberation times T_R of 0.1, 0.2, 0.5, 1.0, and 2.0 s were convoluted into the original speech.

For noisy reverberant emotional speech, a reverberant speech was created by convolving three types of room-impulse responses with T_R of 0.5, 1.0, and 2.0 s into the original speech. Then, five types of constant noise (white Gaussian noise) with SNRs of 20, 10, 5, 0, and -5 dB were added to the reverberant speech. From the combination of the above cases, there were a total of 15 reverberation conditions.

After adding noise and reverberation, the experimental stimuli of NVS were created. NVS is speech obtained by driving the TAE information using band-limited random noise as a carrier signal. Figure 2 shows the generation of the NVS stimuli. As the input signal, noisy reverberant emotional speech was divided into several frequency bands by using an auditory filterbank that simulates human frequency selectivity. The 6^{th}-order Butterworth IIR band-pass filterbank was used as the auditory filterbank. The relationship between $\mathrm{ERB_N}$-number and acoustic frequency is defined as

$$\mathrm{ERB_N{-}number} = 21.4\log_{10}\left(\frac{4.37f}{1000} + 1\right), \qquad (13)$$

where f is the frequency in Hz, and the subscript N indicates the characteristics of normal hearing. The signal was then constructed at the boundary frequencies of the BPFs, BPFs were defined as $\mathrm{ERB_N}$-number from 3 to 35 Cam with bandwidths of 2 $\mathrm{ERB_N}$, and the number of channels of the BPFs was 16.

In each frequency band, the TAE of the signal was extracted using the Hilbert transform and a 2^{nd}-order Butterworth IIR LPF with a cut-off frequency of 64 Hz.

The TAE in each channel was then served with the band-limited noise generated by band-pass filtering white Gaussian noise at the same boundary frequency. All amplitude-modulated noise was summed to generate the NVS stimulus. The sampling frequency for stimulus creation was unified at 20 kHz.

C. Speech emotion recognition experiment setting

Since we used a small dataset [17] to evaluate the effectiveness of the MSFs on emotion recognition, we used a SVM as it effectively captures the data characteristics in a less-data situation. In particular, we used the SVM as the classifier to

conduct a five-class classification. The five classes correspond to the five emotion labels in the Fujitsu dataset, which are happy, sad, natural, cold anger, and hot anger. The input to the model is the speech feature including the MSFs and other feature sets as baselines, which are described below. We investigated each noisy reverberant environment described in Section III. B. The input features were extracted from each corresponding speech environment.

We aimed to comprehensively evaluate the effect of each single MSF and combinations of MSFs. Accordingly, we divided the input of MSFs into two groups: a single MSF and combination of MSFs. For the single MSF, we used each single MSF as one input vector to the SVM. The input is a one-dimensional vector accordingly. We evaluated each of the ten MSFs in each speech environment. For the combination of MSFs, we used each combination (contains several single MSF) as the input to the SVM, so that the input's dimension would be from 2 to 10 in accordance with the features' combination. As there are 1013 combinations from 10 MSFs, we evaluated each of the 1013 combinations in each speech environment.

We used two widely used feature sets as baseline features to evaluate the effectiveness of the MSFs in each speech environment:

The InterSpeech 2009 emotion feature set (IS09) [20] was proposed for the emotion recognition challenge on InterSpeech 2009. This feature set contains 32 types of low-level descriptors (LLD), each LLD has 12 functional features. The features include fundamental frequency (F0), zero cross rate, and MFCCs 1-12. The total dimension of this feature set is 384.

The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [21] was proposed as a minimalistic feature set to provide a basic standard acoustic parameter for various areas of automatic voice analysis. The eGeMAPS contains frequency, energy, amplitude, and spectral features. The dimension of the eGeMAPS is 88.

To reduce the effect of data distribution in the evaluation, we used a five-fold training strategy. We split the dataset into five folds, each time we used four of the folds as a training set to train the model, and the remaining fold was used as the test set to evaluate the model's performance. We trained five SVMs for each feature input in one noisy reverberant environment. The average performance of the five SVMs was used to evaluate the input feature's performance in the noisy and reverberant environments. Unweighted accuracy was used as the evaluation metric.

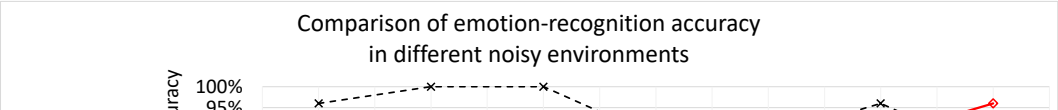
IV. RESULTS

In all figures in this section, the black line indicates the emotion-recognition accuracy of IS09, gray line indicates that of eGeMAPS, green line indicates the highest emotion-recognition accuracy of a single MSF among the ten MSFs, and red line indicates the MSFs combination with the highest emotion-recognition accuracy.

Figure 3 shows the results of the comparison of emotion-recognition accuracy in different noisy environments. The IS09 feature set performed better than eGeMAPS in all noisy environments. The best combination of MSFs improved the emotion-recognition accuracy by 4, 4, and 12% compared with IS09 in the noisy environments of SNR of 10, 5, and -5 dB, respectively. Although the best single MSF did not perform better than the IS09, the best single MSF improved the emotion-recognition accuracy by 4, 8, and 8% compared with eGeMAPS in the noisy environments of SNR of 5, 0, and -5 dB, respectively. These results indicate that the MSFs contribute to SER in noisy environments. The mean and variance of the best combination of MSFs among all noisy environments were 90.67 and 0.11%. While the mean and variance of IS09 were 92.67 and 0.47%, those of eGeMAPS were 80.67 and 0.22%. The performance of the best combination of MSFs was more stable than those of IS09 and eGeMAPS. These results indicate that MSFs are robust against noise conditions.

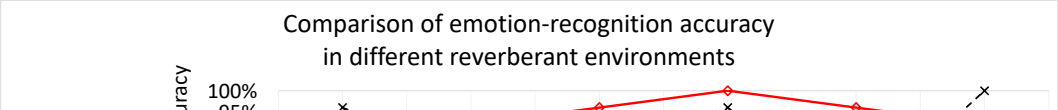
Figure 4 shows the results of the comparison of emotion accuracy in different reverberant environments. The IS09 feature set performed better than eGeMAPS in all reverberant environments. The best combination of MSFs improved the emotion-recognition accuracy by 8, 12, 4, and 24% compared with ISO9 in the reverberant environments of T_{R} of 0.1, 0.2, 0.5, and 1.0 s, respectively. The best single MSF improved the emotion-recognition accuracy by 8 and 8% compared with eGeMAPS in the reverberant environments of T_{R} of 1.0 and 2.0 s, respectively. The mean and variance of the best combination of MSFs among all reverberation environments were 95.20 and 0.11%. While the mean and variance of IS09 were 87.20 and 1.23%, those of eGeMAPS were 71.20 and 0.27%. The performance of the best combination of MSFs was more stable than those of IS09 and eGeMAPS. These results indicate that MSFs are robust against reverberation conditions.

Figures 5-9 show the results of the comparison of emotion accuracy in different noisy reverberant environments. Except for the extremely harsh noisy reverberant environment ($\mathrm{SNR} = -5$ dB and $T_{\mathrm{R}} = 2.0$ s), IS09 performed better performance than eGeMAPS in all noisy reverberant environments. The best combination of MSFs improved the emotion-recognition accuracy by 8, 20, 8, 28, 8, and 8% compared with ISO9 in the noisy reverberant environments of $\mathrm{SNR} = 5$ dB and $T_{\mathrm{R}} = 0.5$ s, $\mathrm{SNR} = 0$ dB and $T_{\mathrm{R}} = 1.0$ s, $\mathrm{SNR} = 0$ dB and $T_{\mathrm{R}} = 2.0$ s, $\mathrm{SNR} = -5$ dB and $T_{\mathrm{R}} = 0.5$ s, and $\mathrm{SNR} = -5$ dB and $T_{\mathrm{R}} = 1.0$ s, respectively. The best combination of MSFs improved the emotion-recognition accuracy by 8% compared with eGeMAPS in the noisy reverberant environment of $\mathrm{SNR} = -5$ dB and $T_{\mathrm{R}} = 2.0$ s. The mean and variance of the best combination of MSFs among all noise reverberation environments were 90.93 and 0.15%. While the mean and variance of IS09 were 87.73 and 1.36%, those of eGeMAPS were 71.47 and 1.19%. The performance of the best combination of MSFs was more stable than IS09 and eGeMAPS. These results indicate that MSFs are robust against noisy reverberant environments.



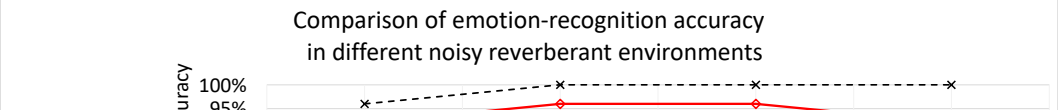
	Original NVS	SNR 20 dB	SNR 15 dB	SNR 10 dB	SNR 5 dB	SNR 0 dB	SNR-5 dB
IS09	96%	100%	100%	88%	88%	96%	84%
eGeMAPS	76%	84%	88%	76%	80%	76%	80%
Best single MSF	68%	72%	80%	76%	84%	84%	88%
Best combination of MSFs	92%	88%	88%	92%	92%	88%	96%

Fig. 3. Results of comparing emotion accuracy in different noisy environments



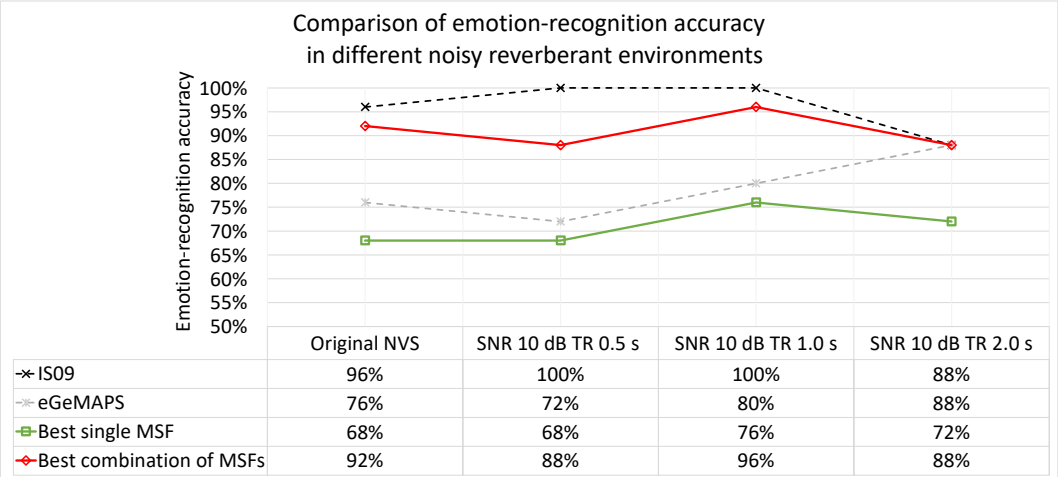
	Original NVS	TR 0.1 s	TR 0.2 s	TR 0.5 s	TR 1.0 s	TR 2.0s
IS09	96%	84%	84%	96%	72%	100%
eGeMAPS	76%	68%	76%	76%	64%	72%
Best single MSF	68%	52%	76%	72%	72%	80%
Best combination of MSFs	92%	92%	96%	100%	96%	92%

Fig. 4. Results of comparing emotion accuracy in different reverberant environments



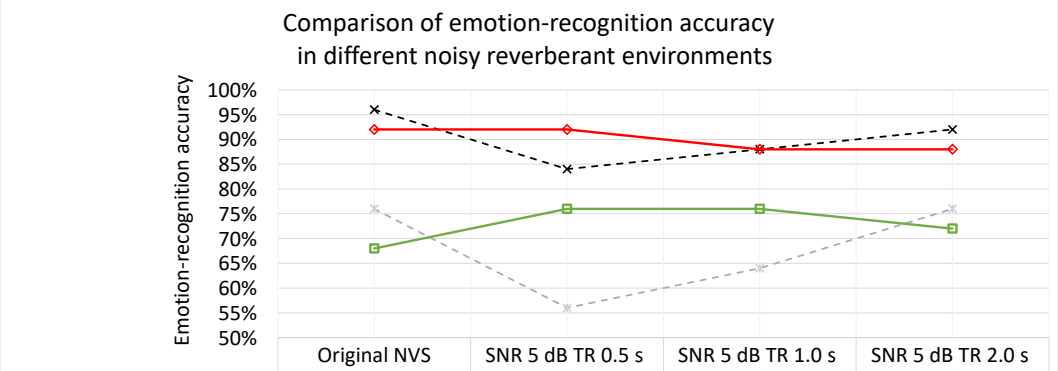
	Original NVS	SNR 20 dB TR 0.5 s	SNR 20 dB TR 1.0 s	SNR 20 dB TR 2.0 s
IS09	96%	100%	100%	100%
eGeMAPS	76%	84%	80%	76%
Best single MSF	68%	76%	64%	68%
Best combination of MSFs	92%	96%	96%	92%

Fig. 5. Results of comparing emotion accuracy in different noisy reverberant environments (noise condition was SNR= 20 dB, reverberation condition changed)



	Original NVS	SNR 10 dB TR 0.5 s	SNR 10 dB TR 1.0 s	SNR 10 dB TR 2.0 s
IS09	96%	100%	100%	88%
eGeMAPS	76%	72%	80%	88%
Best single MSF	68%	68%	76%	72%
Best combination of MSFs	92%	88%	96%	88%

Fig. 6. Results of comparing emotion accuracy in different noisy reverberant environments (noise condition was SNR= 10 dB, reverberation condition changed)



	Original NVS	SNR 5 dB TR 0.5 s	SNR 5 dB TR 1.0 s	SNR 5 dB TR 2.0 s
IS09	96%	84%	88%	92%
eGeMAPS	76%	56%	64%	76%
Best single MSF	68%	76%	76%	72%
Best combination of MSFs	92%	92%	88%	88%

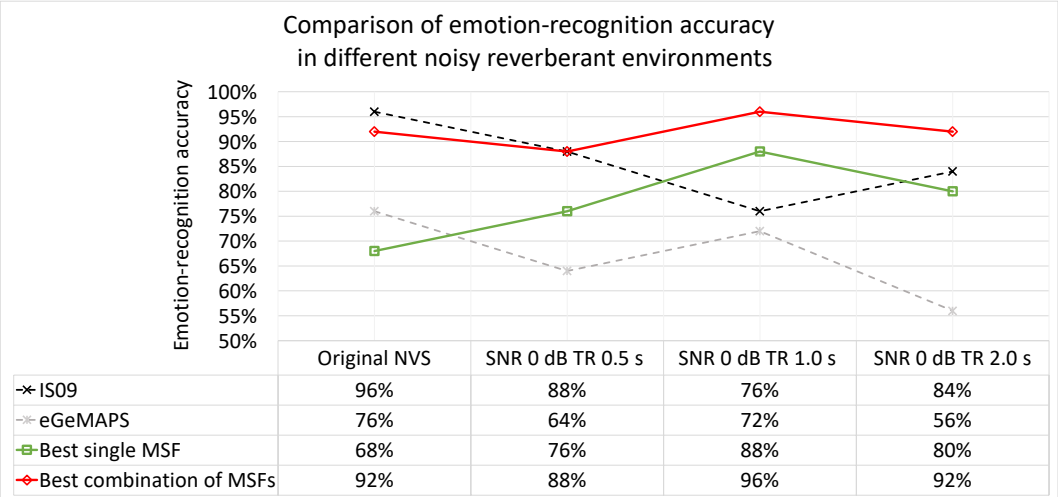
Fig. 7. Results of comparing emotion accuracy in different noisy reverberant environments (noise condition was SNR= 5 dB, reverberation condition changed)

In summary, the experimental results indicate that using MSFs improved the recognition accuracy in 13 of the 26 environments compared with the best baseline feature set by an average improvement of 11.38%. These results also indicate that, by describing the TAE characteristics of the speech signals, the MSFs are robust against noise and reverberation for SER.

V. DISCUSSION

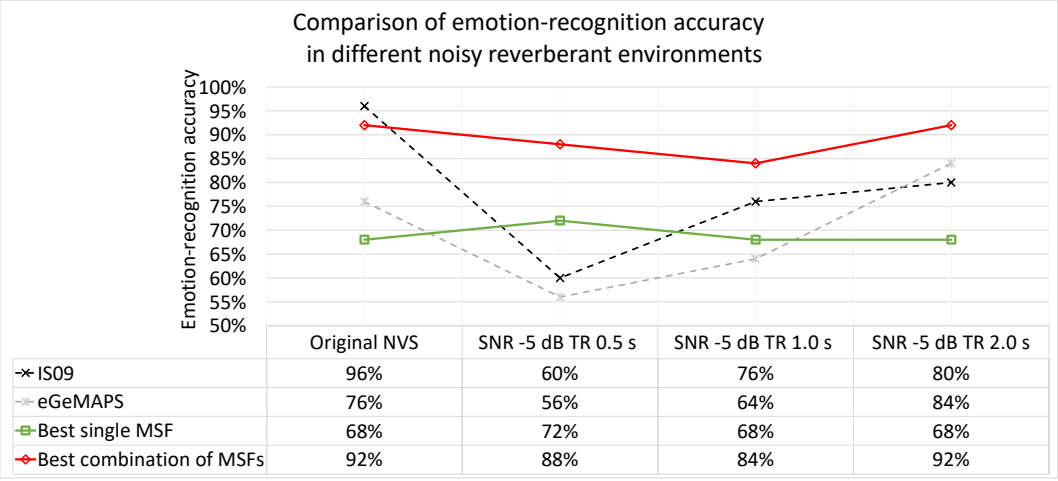
As mentioned in previous sections, using MSFs has the potential to improve emotion-recognition rates in noisy reverberant environments. The results are consistent with the expectation that MSFs are robust against noisy and reverberant environments for SER. This is assumed to due to the extraction method of MSFs. MSFs were proposed on the basis of the auditory modulation filterbank in the human auditory system [17]. Thus, MSFs can reflect the characteristics of the process of human vocal-emotion recognition while having noise-reverberation-robustness similar to that of human vocal-emotion recognition.

As the above results indicate, except for the extremely harsh noisy reverberant environment ($\mathrm{SNR} = -5$ dB and $T_{\mathrm{R}} = 2.0$ s), the best combination of MSFs always achieved higher recognition accuracy than the best single MSF. This can be explained by the fact that different MSFs represent different higher-order statistics in the processing of emotional speech signals, and these high-order statistics play different roles in different emotion-recognition processes. For example, a breathy voice, such as sad speech, is produced when the vocal fold motion is not broad enough to close the glottis completely. This phenomenon makes a breathy voice have a spectrum with a strong slope [22]. Consequently, the slope-related modulation spectral tilts (MSTL_m and MSTL_k) are important in sadness recognition. When we combine the MSFs, different MSFs complement each other; thus, combinations of MSFs have advantages during average emotional recognition



	Original NVS	SNR 0 dB TR 0.5 s	SNR 0 dB TR 1.0 s	SNR 0 dB TR 2.0 s
IS09	96%	88%	76%	84%
eGeMAPS	76%	64%	72%	56%
Best single MSF	68%	76%	88%	80%
Best combination of MSFs	92%	88%	96%	92%

Fig. 8. Results of comparing emotion accuracy in different noisy reverberant environments (noise condition was SNR= 0 dB, reverberation condition changed)



	Original NVS	SNR -5 dB TR 0.5 s	SNR -5 dB TR 1.0 s	SNR -5 dB TR 2.0 s
IS09	96%	60%	76%	80%
eGeMAPS	76%	56%	64%	84%
Best single MSF	68%	72%	68%	68%
Best combination of MSFs	92%	88%	84%	92%

Fig. 9. Results of comparing emotion accuracy in different noisy reverberant environments (noise condition was SNR=-5 dB, reverberation condition changed)

over a single MSF.

VI. CONCLUSIONS

We investigated MSFs as robust features against noise and reverberation for SER. We used NVS as speech data and an SVM as the classifier to carry out emotion recognition. The experimental results indicate that, compared with two widely used feature sets, IS09 and eGeMAPS, using MSFs (containing the best single MSF and best combination of MSFs) improved the recognition accuracy in 13 of the 26 environments, with an average improvement of 11.38%. These results also indicated that the MSFs contribute to SER and are robust against noise and reverberation.

For future work, we will further analyze the specific components contained in the combination of MSFs. If an optimal combination achieves the highest recognition rate many times in all noisy reverberant environments, then this combination can be considered an important combination of MSFs for SER. By analyzing the important combinations of MSFs, a robust and better-performing SER model can be designed.

ACKNOWLEDGMENT

This research was supported by JST SPRING (Grant Number JPMJSP2102), a Grant-in-Aid for Scientific Research (B) (Grant Number 21H03463), SCOPE Program of Ministry of Internal Affairs and Communications (Grant Number 201605002), Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Numbers 22H04860 and 22H00536), and JST AIP Trilateral AI Research, Japan (Grant Number JPMJCR20G6).

REFERENCES

[1] Kanedera, N., Arai, T., Hermansky, H., Pavel, M. , “On the relative importance of various components of the modulation spectrum for automatic speech recognition,” *Speech Communication,* 28(1), 43-55, 1999.

[2] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B. , "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine,* 29(6), 82-97, 2012.

[3] Baker, J. M., Deng, L., Glass, J., Khudanpur, S.. Lee, C. H., Morgan, N., O'Shaughnessy, D. , "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]," *IEEE Signal processing magazine,*26(3), 75-80, 2009.

[4] Nygaard, L. C., Tzeng, C. Y. , " Perceptual integration of linguistic and non-linguistic properties of speech," *The handbook of speech perception,*398-427, 2021

[5] Akagi, M. ,"Analysis of production and perception characteristics of non-linguistic information in speech and its application to inter-language communications," *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference,* pp. 513-519, Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009.

[6] Likitha, M.S., Gupta, S.R.R, Hasitha, K., Raju, A. U. , "Speech based human emotion recognition using MFCC.," *2017 international conference on wireless communications, signal processing and networking (WiSPNET),* IEEE, 2017.

[7] Dahake, P.P., Shaw, K., Malathi P. ,"Speaker dependent speech emotion recognition using MFCC and Support Vector Machine," *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT).*, IEEE, 2016.

[8] Heracleous, P., Yasuda, K., Sugaya, F., Yoneyama, A., Hashimoto, M., "Speech emotion recognition in noisy and reverberant environments," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII),* pp. 262-266, IEEE, 2017.

[9] Bashirpour, M., Geravanchizadeh, M., "Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments," *EURASIP Journal on Audio, Speech, and Music Processing,* 1, 1-13, 2018.

[10] Chenchah, F., Lachiri, Z., "Speech emotion recognition in noisy environment," *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP),* IEEE, 2016.

[11] Zhu, Z., Kawamura, M., Unoki, M. , "Study on the perception of nonlinguistic information of noise-vocoded speech under noise and/or reverberation conditions," *Acoustical Society and Technology,* 2022 (in press)

[12] Zeng, F.G., Rebscher, S., Harrison, W., Sun, X., Feng, H., "Cochlear implants: system design, integration, and evaluation," *IEEE reviews in biomedical engineering,* 1, 115-142, 2008.

[13] Xiang, J., Poeppel, D., Simon, J.Z. , "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *The Journal of the Acoustical Society of America,* 133(1), EL7-EL12, 2013.

[14] Tachibana, R.O., Sasaki Y., Riquimaroux H., "Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech," *Acoustical Science and Technology,* 34.4: 263-270,2013.

[15] Xu, L., Bryan, E.P., "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hearing research,* 242.1-2, 132-140, 2008.

[16] Unoki, M., Zhu, Z., "Relationship between contributions of temporal amplitude envelope of speech and modulation transfer function in room acoustics to perception of noise-vocoded speech," *Acoustical Science and Technology,* 41.1, 233-244, 2020.

[17] Zhu, Z., Miyauchi, R., Araki, Y., Unoki, M., "Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech," *Acoustical Society and Technology,* 39(6), 379- 386, 2018.

[18] Moore, B.C.J. , *An Introduction to the Psychology of Hearing,*sixth edition, Brill Academic Pub, 2013.

[19] Schroeder, M. R., "Modulation transfer functions:10 definition and measurement," *Acta Acustica united with Acustica,* 49(3): 179-182, 1981.

[20] Schuller, B., Steidl, S., Batliner, A., "The interspeech 2009 emotion challenge," 2009.

[21] Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C. , Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K. P., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing,* 7(2), 190-202, 2015.

[22] Ishi, C. T., Ishiguro, H., Hagita, N., "Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech," *EURASIP Journal on Audio, Speech, and Music Processing,*1-12, 2010.