2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) | 979-8-3503-0067-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/APSIPAASC58517.2023.10317449

Contribution of modulation spectral features for cross-lingual speech emotion recognition under noisy reverberant conditions

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

***Abstract*—Handling multiple languages under noisy reverberant conditions has become increasingly important for speech emotion recognition (SER). Previous studies found that modulation spectral features (MSFs) are robust to noisy reverberant conditions for SER. However, they mainly focused on specific languages; the universality of MSFs among languages is still unclear. To address this issue, we compared MSFs, hand-crafted features, Wav2Vec2.0-based features, MSFs+hand-crafted features for SER on four languages under 12 noisy reverberant conditions. Intra-lingual results showed that MSFs+hand-crafted features performed best on most conditions of all languages. Inter-lingual results showed that MSFs performed best on most conditions of test languages except training on a tonal language and testing on others. The results demonstrate that MSFs are robust to multilingual SER under noisy reverberant conditions and suggest that MSFs are potentially language-independent features for nontonal languages.**

I. Introduction

Speech-emotion recognition (SER) has become an increasingly important technique in a wide range of fields, including psychology for detecting emotional disorders. In real-world applications, noise and reverberation can have a significant impact on the quality and clarity of speech signals [1]–[5].

To address the challenge of the lack of robustness against noise and reverberation in SER, researchers have proposed methods based on speech signal processing to improve SER. Modern physiological [6] and psychological [7] models suggest that the human auditory system contains an auditory filterbank that decomposes speech signals into channel signals, such as the temporal amplitude envelope (TAE) and temporal fine structure. Several studies have proved that the TAE and its modulation cues play an important role in speech recognition [8]–[10]. Based on these findings, a previous study proposed modulation spectral features (MSFs) that were extracted based on the modulation analysis processing of the TAE and showed that MSFs are important cues for human emotion recognition [11]. Furthermore, MSFs were shown to be more robust on SER under noisy and reverberant conditions compared to widely used feature sets IS09 and eGeMAPS [12], [13].

So far, studies on MSFs were mainly focused on specific languages, such as Japanese [12], [13] and German [14]. However, handling multiple languages is also important for SER [15], [16]. Whether or not MSFs retain their robustness in multilingual SER and remain unaffected by noise and reverberation is still unclear. On the other hand, inspired by human perception, researchers have recently proposed to use spectral features with the three-layer model for multilingual emotion recognition tasks [17]–[19]. This suggests that spectral features inspired by human perception, such as MSFs, may reflect the essential characteristic of emotions under noisy reverberant conditions among languages.

Accordingly, this study investigated the potential universality of MSFs in intra-lingual and inter-lingual SER under noisy and reverberant conditions. To do so, we used support vector machine (SVM) to compare the SER performance of MSFs, hand-crafted feature sets, MSFs + hand-crafted feature sets, and Wav2Vec2.0-based features on different languages under noisy reverberant conditions. Specially, we investigated intra-lingual and inter-lingual SER performances on recognizing four common emotions (happy, neutral, sad, and angry) in four languages (Japanese, German, English and Chinese) under 12 noisy reverberant conditions. To the best of our knowledge, we are the first to investigate the robustness of MSFs in multilingual and cross-lingual SER under noisy reverberant conditions. The findings of this study will not only advance the understanding of the importance of MSFs in the speech perception mechanism, but also contribute to the development of robust SER systems, such as speech-based human-machine interaction under noisy reverberant conditions.

II. Modulation spectral features

We utilized the same MSFs that were shown to be robust to noise and reverberation in previous studies [11]–[13]. To extract modulation spectral cues such as MSFs of modulation spectrogram, it is necessary to calculate the modulation spectrogram using the modulation filterbank first. Figure 1 shows the auditory-based process used in this study to calculate the MSFs of modulation spectrograms.

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

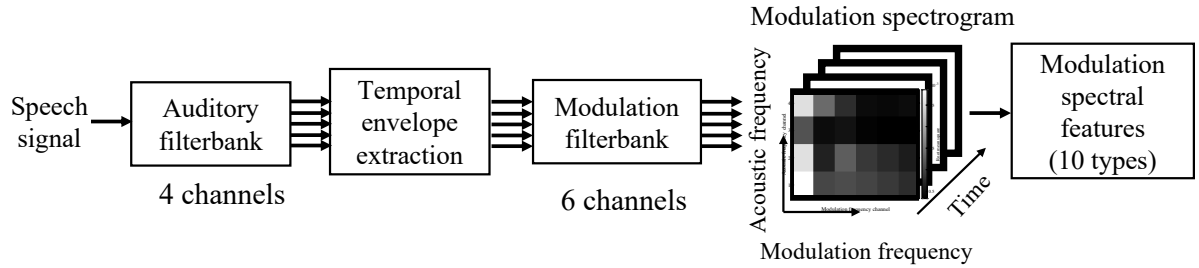

Fig. 1. Process of extracting MSFs [13].

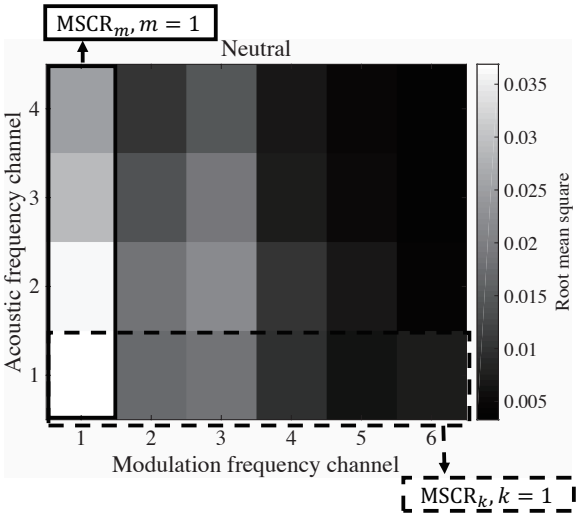

Fig. 2. Calculation example of MSCR$_m$ and MSCR$_k$ [11].

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, 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) 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 $s(n)$.

Then, we calculated the high-order statistics of the three-dimensional modulation spectrograms (time, acoustic frequency, modulation frequency) in the acoustic frequency direction and modulation frequency direction as MSFs. The ten types of MSFs are the modulation spectral features in the acoustic frequency domain (the subscript is m) and in the modulation frequency domain(the subscript is k): the modulation spectral centroid (MSCR$_{m/k}$), modulation spectral spread (MSSP$_{m/k}$), modulation spectral skewness (MSSK$_{m/k}$) and modulation spectral kurtosis (MSKT$_{m/k}$), which are defined as follows.

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

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

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

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

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

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

$$\mathrm{MSSK_k} = \frac{\sum_{m=1}^{M}[m - MSCR_k]^3 \bar{E}_{k,m}}{\sum_{m=1}^{M} \bar{E}_{k,m}} \quad (11)$$

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

The final two MSFs, the modulation spectral tilts ($\mathrm{MSTL}_{m/k}$) are the linear regression coefficients obtained by fitting the first-degree polynomial to the modulation spectrograms. Figure 2 shows an example of calculating MSCR_m and MSCR_k.

III. Noisy reverberant conditions

In this study, we used the same method as in previous studies [11]–[13] to add noise and reverberation to speech to simulate noisy, reverberant, and noisy reverberant conditions.

For noisy conditions, we added white Gaussian noise and the adjusted noise to the speech at SNR levels of 20, 10, and -5 dB. The noise conditions were three in total. For reverberant conditions, we used the convolution method and a statistical room impulse response (Schroeder model). We convoluted the original speech with two types of room-impulse responses, with reverberation times (T_{R}) of 1.0 and 2.0 s, respectively. The reverberation conditions were two in total. For noisy reverberant conditions, we used the combination of the above three noise conditions and two reverberation conditions. The noisy reverberant conditions were six in total. With the clean condition (original speech), we conducted experiments under 12 conditions in total.

IV. Experiments

A. Datasets

We used datasets of Japanese, German, Chinese, and English in this study. We chose happy, neutral, sad, and angry emotions in these datasets, because these emotions are common emotion categories that guarantee the emotion categories in each dataset are the same for experiments. Table 1 shows the statistics of utterances of emotions in each dataset.

Fujitsu dataset: The Fujitsu Japanese Emotional Speech Database is a Japanese dataset that was used for emotion recognition [11]–[13]. **Berlin EmoDB dataset**: This is a German dataset, which has also been widely used in previous research [18]–[20]. For convenience, we will refer to Berlin EmoDB as Berlin for short. **IEMOCAP dataset**: IEMOCAP is an interactive emotional dyadic motion capture database in English [21], [22]. For emotion categories, we combine the emotions excited and happy together as the emotion happy, as was done in previous studies [23], [24]. **CASIA dataset**: The CASIA emotional speech dataset [18], [19], [22] is a Mandarin emotional speech dataset. This dataset does not contain the emotion of happy, but it has joy, a closely related emotion. Accordingly, we use joy as the emotion happy for this dataset.

TABLE I
Statistics of utterances of emotions in each dataset

Dataset	Happy	Neutral	Sad	Angry	Total
Fujitsu	20	20	20	20	80
Berlin	71	79	62	127	339
IEMOCAP	1636	1708	1084	1130	5531
CASIA	1600	400	1599	1597	5196

B. Baseline feature sets

Hand-crafted feature sets: InterSpeech2009 (IS09) [25] and eGeMAPS [26] are two widely used hand-crafted feature sets in the emotion recognition area. We used two feature sets separately in our experiments.

Wav2Vec2.0-based feature: Wav2Vec2.0 [27] is a self-supervised learning pre-trained model for representing speech. We used Wav2Vec2.0-large-xlsr models that were fine-tuned on each language to make fair comparisons. These models were obtained from jonatasgrosman/Wav2Vec2.0-large-xlsr-53-[Japanese, German, English, Chinese] from HuggingFace. To make a fair comparison with MSFs, we did not fine-tune Wav2Vec2.0 models on SER tasks but froze those models to use them as feature extractors only. For feature extractions, as Wav2Vec2.0's output is a sequence of vectors corresponding to frames sequence, we refered to previous studies [28], [29] to obtain fixed-length features for each speech. Specifically, we computed the mean of the last layer, the sum of the last layer, the mean of all layers, and the sum of all layers as four different features. Each feature is a 1024-dimension vector.

C. MSFs

In this study, we investigated the performance of each of the ten types of MSFs and the four combinations of MSFs. Combination features include the combination of all types of MSFs, the combination of acoustic-frequency domain MSFs, the combination of modulation-frequency domain MSFs, and the modulation spectral tilts (MSKT_k and MSTL_k). The last two features have been demonstrated to be robust to all daily noise and reverberation conditions in a previous study [13]. We concatenated corresponding MSFs together to obtain the combination of features.

Moreover, since MSFs are hand-crafted features, we also used MSFs + hand-crafted feature sets to investigate whether or not MSFs are complementary to conventional feature sets. Specifically, we concatenated all types of MSFs with IS09 and all types of MSFs with eGeMAPS as two kinds of MSFs + hand-crafted features.

D. Intra-lingual and inter-lingual experiment setting

We used SVM as a classifier to conduct experiment recognition experiments, which has also been widely used in many studies related to SER (as examples, see [30] and [31]). The input was the feature vector of MSFs, MSFs + hand crafted, hand crafted, and Wav2Vec2.0-based features. The output was the emotion from the four emotion classes.

For the intra-lingual experiment, we performed five-fold experiments to comprehensively evaluate the performance. Specifically, we split the data of each emotion into a five-fold average for each dataset. Then, we performed five experiments by using four of the folds as the training set and the remaining

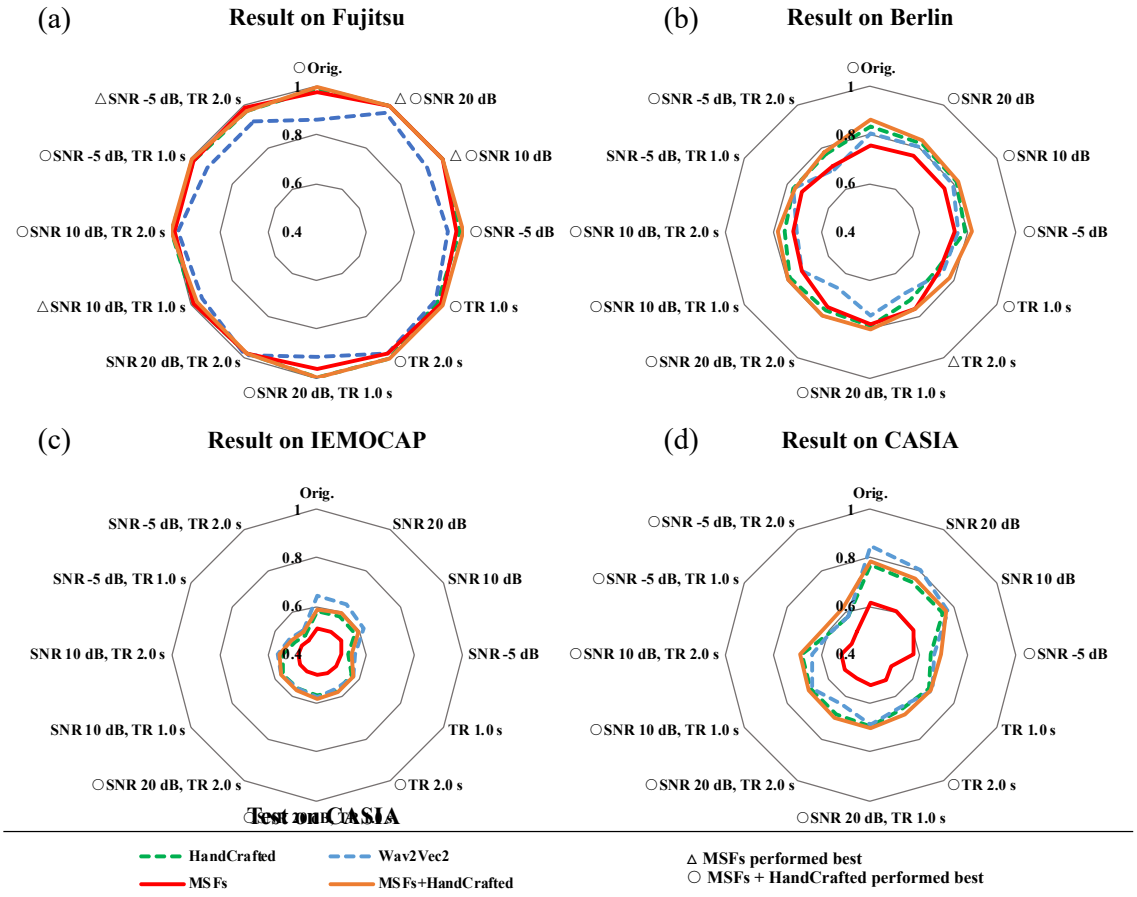

Fig. 3. Result of intra-lingual SER on four datasets: (a) result of train on Fujitsu and test on Fujitsu, (b) result of train on Berlin and test on Berlin, (c) result of train on IEMOCAP and test on IEMOCAP, and (d) result of train on CASIA and test on CASIA.

one as the test set. Moreover, we splited 80% of the training set for training and the remaining 20% for validation. After using each fold as the test set, we used the average performance of the five experiments for evaluation.

For the inter-lingual experiment, we trained the model on one dataset and tested it on the other three datasets separately. Specifically, we used 80% of each emotion's data from the training dataset as the training set and the remaining 20% of each emotion's data for validation. All data from test datasets were used for the test. We conducted experiments by training the model on each dataset and testing it on other datasets to evaluate the performance.

We used weighted accuracy that considers the imbalance of each emotion as the evaluation metric.

V. Results and Discussion

Figure 3 and 4 show radar charts of intra-lingual and inter-lingual results. Different axes represent the original speech and 12 noise reverberant conditions; values on axes represent the weighted accuracy. To show comparisons between four feature sets clearly, and due to space limitations, we only show the best result in categories of hand-crafted feature sets, Wav2Vec2.0-based features, MSFs, and MSFs + hand-crafted feature sets. For convenience, we hereafter refer to these categories as hand-crafted, Wav2Vec2.0-based, MSFs, and MSFs + hand-crafted. In Figure 3 and 4, different colored lines represent different feature categories: the green dashed line, blue dashed line, red line, and orange line indicate the highest weighted accuracy from hand-crafted, Wav2ec2-based, MSFs, and MSFs + hand-crafted, respectively. Conditions with △ indicate conditions where MSFs performed the best. Conditions with ◯ indicate conditions where MSFs + hand-crafted performed the best.

As seen in Fig. 3, MSFs + hand-crafted performed best on 9 of 12 conditions on Fujitsu, 10 of 12 conditions on Berlin, and 8 of 12 conditions on CASIA, with average improvements from the best baseline of 0.69%, 2.03%, and 1.85% on Fujitsu, Berlin, and CASIA, respectively. The reason MSFs + hand-crafted improved relatively less is because the best baselines and MSFs + hand-crafted have 100% weighted accuracy on many conditions in the Fujitsu dataset. On the other hand, MSFs + hand-crafted performed best on 3 of 12 conditions on IEMOCAP, while Wav2Vec2.0-based performed best on more conditions in IEMOCAP. The reason could be that Wav2Vec2.0 was originally pre-trained in English, and it captures characteristics of emotions in English well. However,

Fig. 4. Result of inter-lingual SER that train the model on a single language and test on other languages: (a) result of train on Fujitsu and test on other datasets, (b) result of train on BERLIN and test on other datasets, (c) result of train on IEMOCAP and test on other datasets, and (d) result of train on CASIA and test on other datasets.

although Wav2Vec2.0 models were fine-tuned on languages other than English, MSFs + hand-crafted performed better than Wav2Vec2.0-based on most conditions of other languages. Meanwhile, Wav2Vec2.0 is a large model that needs more computational resources, while MSFs + hand-crafted can be effectively obtained. These results demonstrated that MSFs are complementary with conventional hand-crafted features and are effective features that are robust in noisy reverberant conditions in different languages.

As seen in Fig. 4, MSFs performed best on most of the conditions when training on Fujitsu and testing on other datasets, training on Berlin and testing on Fujitsu and CASIA, and training on IEMOCAP and testing on other datasets, training on CASIA and testing on Fujitsu. MSFs + hand-crafted performed best on most conditions when training on Berlin and testing on IEMOCAP. These results demonstrated that MSFs are robust in emotion recognition in noisy reverberant conditions across languages. On the other hand, MSFs or MSFs

+ hand-crafted performed best on relatively fewer conditions when training on CASIA and testing on Berlin and IEMOCAP. Since Chinese was the only tonal language in our experiments, the emotion perception in Chinese may be influenced by tone and other reasons. These results suggested that MSFs potentially describe common characteristics of emotions under noisy reverberant conditions among non-tonal languages.

For the MSFs combination that performed best in the results, among all 48 conditions (12 conditions $\times$ 4 datasets) in multilingual SER, features containing MSFs performed best on 33 conditions, and among them, 30 combinations are MSFs + hand-crafted features. On the other hand, among all 144 conditions (12 conditions $\times$ 4 datasets $\times$ 3 cross-lingual) in cross-lingual SER, features containing MSFs performed best on 120 conditions, and among them, 29 of the best performing features are $\mathrm{MSSK_m}$. Therefore, $\mathrm{MSSK_m}$ is considered the universal feature in cross-lingual SER.

VI. Conclusion

This study investigated the robustness of MSFs in multilingual and cross-lingual SER under noisy reverberant conditions. We utilized SVM to compare the SER performance for recognizing four common emotions. These comparisons were conducted on intra-lingual and inter-lingual SER in four languages under 12 noisy reverberant conditions. The results demonstrate that MSFs are complementary with conventional hand-crafted features and they are robust under noisy reverberant conditions in different languages. The results also suggest that MSFs potentially describe common characteristics of emotions in noisy reverberant conditions among non-tonal languages.

For future work, it would be interesting to investigate the applicability of MSFs in recognizing emotions in other tonal languages to further examine the language-dependency of these features. Additionally, it would be worthwhile to explore the use of deep learning techniques to further enhance the performance of SER systems utilizing MSFs.

Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research (B) (21H03463), a Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233), the Japan Society for the Promotion of Science (JSPS) KAKENHI (No. 22K21304 and 22H00536), JST AIP Trilateral AI Research, Japan (No. JPMJCR20G6) and JST Moonshot R&D program (JPMJMS2237).

References

[1] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 177–188, 2018.

[2] S. R. Kshirsagar and T. H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1892–1905, 2022.

[3] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *2016 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2016, pp. 5795–5799.

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

[5] A. R. Avila, J. Monteiro, D. O'Shaughneussy, and T. H. Falk, "Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks," in *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, IEEE, 2017, pp. 360–365.

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

[7] J. Xiang, D. Poeppel, and J. Z. Simon, "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, EL7–EL12, 2013.

[8] R. O. Tachibana, Y. Sasaki, and H. Riquimaroux, "Relative contributions of spectral and temporal resolutions to the perception of syllables, words, and sentences in noise-vocoded speech," *Acoustical Science and Technology*, vol. 34, no. 4, pp. 263–270, 2013.

[9] L. Xu and B. E. Pfingst, "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hearing research*, vol. 242, no. 1-2, pp. 132–140, 2008.

[10] M. Unoki and Z. Zhu, "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*, vol. 41, no. 1, pp. 233–244, 2020.

[11] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech," *Acoustical Science and Technology*, vol. 39, no. 6, pp. 379–386, 2018.

[12] T. Guo, S. Li, M. Unoki, and S. Okada, "Investigation of noise-reverberation-robustness of modulation spectral features for speech-emotion recognition," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2022, pp. 39–46.

[13] T. Guo, Z. Zhu, S. Kidani, and M. Unoki, "Contribution of common modulation spectral features to vocal-emotion recognition of noise-vocoded speech in noisy reverberant environments," *Applied Sciences*, vol. 12, no. 19, p. 9979, 2022.

[14] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication*, vol. 53, no. 5, pp. 768–785, 2011.

[15] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, pp. 1–10, 2021.

[16] M. Neumann *et al.*, "Cross-lingual and multilingual speech emotion recognition on english and french," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5769–5773.

[17] C.-F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[18] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Communication*, vol. 110, pp. 1–12, 2019.

[19] X. Li and M. Akagi, "A three-layer emotion perception model for valence and arousal-based detection from multilingual speech," 2018.

[20] X. Mao, L. Chen, and L. Fu, "Multi-level speech emotion recognition based on hmm and ann," in *2009 WRI World congress on computer science and information engineering*, IEEE, vol. 7, 2009, pp. 225–229.

[21] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[22] S. Mao, D. Tao, G. Zhang, P. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6715–6719.

[23] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[24] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning.," in *Interspeech*, vol. 2021, 2021, pp. 4508–4512.

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

[26] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[28] O. Mohamed and S. A. Aly, "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset," *arXiv preprint arXiv:2110.04425*, 2021.

[29] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[30] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proceedings of 2011 international conference on electronic & mechanical engineering and information technology*, IEEE, vol. 2, 2011, pp. 621–625.

[31] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.