Speech emotion recognition using auditory spectrogram and cepstral features

1st Shujie Zhao
Northwestern Polytechnical University
Xi'an, Shaanxi 710072, China
zhaoshujie1126@163.com

2nd Yan Yang
Northwestern Polytechnical University
Xi'an, Shaanxi 710072, China
y.yang@nwpu.edu.cn

2nd Israel Cohen
Technion–Israel Institute of Technology
Haifa 3200003, Israel
icohen@ee.technion.ac.il

3rd Jingdong Chen
Northwestern Polytechnical University
Xi'an, Shaanxi 710072, China
jingdongchen@ieee.org

3rd Lijun Zhang
Northwestern Polytechnical University
Xi'an, Shaanxi 710072, China
zhanglj7385@nwpu.edu.cn

***Abstract*—A systematic comparison on the impact of environmental noises on key acoustic features is critical in order to transfer SER systems into real world applications. In this study, we investigate the noise-tolerance of different acoustic features in distinguishing various emotions by comparing the classification performance in SER on clean speech signals and noisy speech signals. We novelly extracted the spectrum and cepstral parameters based on human auditory characteristics and developed machine learning algorithms to classify four types of emotions using these features. Experimental results across the clean and noisy data show that auditory spectrogram-based features can achieve a higher recognition accuracy than cepstral features under lower SNRs, but a lower accuracy rate under higher SNRs. Gammatone Filter Cepstral Coefficients (GFCCs) outperformed above 15dB among the whole extracted features on EmoDB under all four kinds of noise conditions. These results display some potential compensation relationships between auditory spectrogram-based features and cepstral features for SER with better noise robustness in real world applications.**

***Index Terms*—Emotion recognition, speech signals, machine learning, pattern recognition, feature extraction, noise**

I. Introduction

Emotional aspect of speech is an important factor in human communication. Through a correctly understanding of people's emotions behind their speech, human beings can achieve effective interpersonal communication. Emotion recognition via speech focuses on automatically identifying the affective state of a person from speech and has many applications, e.g., detecting potential problematic points causing anger and frustration in call centers [1], recognizing stress response to a stimuli (questions) in lie detection [2], detecting the uncertainty and correctness for each student turn in spoken dialogue computer tutors [3]. In natural environments, the desired signals for SER usually coexist with background noises. Thus, algorithms developed on clean speech signals will have incorrect SER and reduction in recognition accuracy in real-world noisy condition. Taking signal acquisition in typical indoor environment for example, the echo, reverberation, interference and additive noise, can all lead to degradation of the quality and reliability of speech related recognition tasks [4]–[8]. It is needed to take the noise robustness issue into consideration for SER in real world applications.

Human ears have a good anti-noise recognition ability. Therefore, in speech recognition domain, many studies are devoted to the auditory characteristics of human ears, and many signal processing approximation methods were proposed to simulate the frequency-domain analysis methods of human ears, so as to establish the voice feature parameter model more in line with the auditory characteristics of human ears [9], [10]. Some emotional speaker recognition results show that auditory features can improve the speech recognition results and enhance the noise robustness of the system [11]–[13]. Moreover, some novel features based on human peripheral hearing system were extracted to increase the robust SER performance in noise and reverberation scenarios, such as the supervised Nonnegative Matrix Factorization (NMF) based features [14]–[16], damped oscillator cepstral coefficients (DOCCs) [17], Teager Energy Cepstrum Coefficient (TECC) based features [18], Cochlear filterbank based features [19], [20], pooling scheme-based modulation spectral features [21].

Nevertheless, extracting the effective emotion information from the raw audio data is still an open challenge, since the explicit and deterministic mapping between emotion state and concrete feature at present does not exist. A systematic understanding of the noise robust feature representation for emotional speech is still missing. In this study, we novelly extracted some spectrum and cepstral parameters based on human auditory characteristics. Also we firstly made a fully comparison of the four kinds of spectrum and cepstral parameters on their SER performances across clean data and artificial additive noise data. Simulation results show that auditory spectrogram-based features have a robuster performance than cepstral features under lower SNRs. While cepstral features are able to gain a higher recognition accuracy than auditory

This project was supported in part by the Israel Science Foundation Grant 576/16, in part by ISF-NSFC joint research program Grants 2514/17 and 61761146001, in part by the National Natural Science Foundation of China(NSFC) under grant no. 6177012290 (F011305).

spectrogram based features under higher SNRs.

II. Data Collection

EmoDB is a German open database, including 10 actors (5 female) and 7 types of emotions (neutral, anger, fear, joy, sadness, disgust, and boredom). The data were gathered in the anechoic chamber with a sampling rate of 48kHz, and then down sampled to 16kHz. In the expert testing phase, the emotional utterances that were rated higher than 80% and non-emotional utterances (i.e. naturalness) higher than 60% were retained. Finally, the database includes 535 utterances [22]. In this paper, we only included the speech with happy (joy), angry, neutral and sad emotions for further analysis.

By overlaying the clean speech signals from EmoDB with Gaussian white noise, pink noise, factory noise (factory2.wav) and vehicle noise (volvo.wav) from Noise-92 database, we then generated four extra artificial additive data sets which simulated acted emotions in the presence of background noise. These data were used to evaluate the recognition performance of different speech features in different noise environments under various levels of signal-to-noise ratios (SNRs) (i.e., $-10-40$dB at a step of 5dB), in comparison with recognition performance without noise.

III. Feature Extraction

A. Mel filter

Mel Frequency Cepstral Coefficients (MFCCs) was proposed by Davis and Mermelstein in 1980 [23]. It is now widely used in automatic speech and speaker recognition. The shape of voice-channel can be shown in the envelope of a speech's short-term power spectrum, and MFCCs is a feature that accurately describes this envelope. MFCCs analysis is on the strength of two major auditory mechanisms: the nonlinear classification principle of the subjective perception frequency domain from human beings [24] and classification of critical bands. Here is the following formula:

$$F_{\text{mel}} = 1125 \log(1 + f/700) \tag{1}$$

where, F_{mel} is the perceptive frequency in Mel, and f is the real frequency in Hz. According to the classification of critical bands, the speech is divided into a series of frequency groups in the perceptive frequency domain to form a filter bank, namely a Mel filter bank.

B. Cochlear filter

On the basis of simulating the basement membrane response of the human ear, the cochlear filter realizes the whole process of sound transmission from the outer ear to the basement membrane through wavelet transform, called auditory transformation, which is defined as [25]

$$T(a, b) = \int_{-\infty}^{+\infty} x(t)\psi_{a,b}(t)dt = x(t) * \psi_{a,b}(t) \tag{2}$$

where $x(t)$ is the speech signal in time domain, $\psi_{a,b}(t)$ is the cochlear filter function, $*$ denotes a convolution operation.

$$\begin{aligned} \psi_{a,b}(t) &= \frac{1}{\sqrt{a}}\psi(\frac{t-b}{a}) \\ &= \frac{1}{\sqrt{a}}(\frac{t-b}{a})^{\alpha}\text{exp}(-2\pi f_L \beta(\frac{t-b}{a}) \\ &\quad \cos(2\pi f_L \beta((t-b)/a + \theta))u(t-b) \end{aligned} \tag{3}$$

where α and β are real numbers greater than zero. They together control the shape and width of $\psi_{a,b}(t)$ in frequency domain. In this study, their values were set as 0.3 and 0.2, since the frequency response curve of cochlear filter bank is closer to the auditory frequency response curve of human ear. $u(t)$ is the unit step function, θ is the initial phase, and b is a time varying real value. a is a scale variable determined by the central frequency f_c and the lowest central frequency f_L of the filter bank. $1/\sqrt{a}$ is a energy normalization factor, which ensures a consistent energy under various a and b values. The hair cell function converts the auditory speech signal into the nerve pulse signal. The hair cell window uses different window lengths to analyze different signals, due to the fact that the nerve pulse signal generated by different frequency signals are not the same. Then the energy information of the acquired signal is transformed into perceived loudness through a nonlinear loudness transformation function. Finally, the Cochlear Filter Cepstral Coefficients (CFCCs) can be gained through the DCT transformation.

C. Gammatone filter

In 1992, Roy Patterson and his colleagues designed a Gammatone filter based on the frequency response in the basement membrane of human ears. By simulating the traveling waves in the basement membrane of the cochlea, the time-domain speech signal was decomposed into a series of frequency band information. The impulse response of the Gammatone filter bank is defined as [13]:

$$g_i(t) = At^{n-1}\text{exp}(-2\pi b_i t)\text{cos}(2\pi f_i + \phi_i)u(t) \tag{4}$$

where, $t \geq 0$, and $1 \leq i \leq N$. A is the filter gain, n is the order, f_i is the center frequency, $u(t)$ denotes the step function, N is the number of filters, and ϕ_i is the initial phase. b_i represents the attenuation factor determining the attenuation speed of the impulse response and can be depicted as

$$b_i = 1.019 b_{\text{ERB}}(f_i) \tag{5}$$

where $b_{ERB}(f_i)$ is the Equivalent Rectangular Bandwidth (ERB) of each filter, which is related to the center frequency of the filter and the critical frequency band of human auditory system. $b_{ERB}(f_i)$ in auditory psychology model is given as

$$b_{\text{ERB}}(f_i) = 24.7(4.37 \times \frac{f_i}{1000} + 1) \tag{6}$$

The central frequency is proportional to the bandwidth on a logarithmic scale, that is, it has a nonlinear frequency characteristic and conforms to the auditory characteristics of human ears.

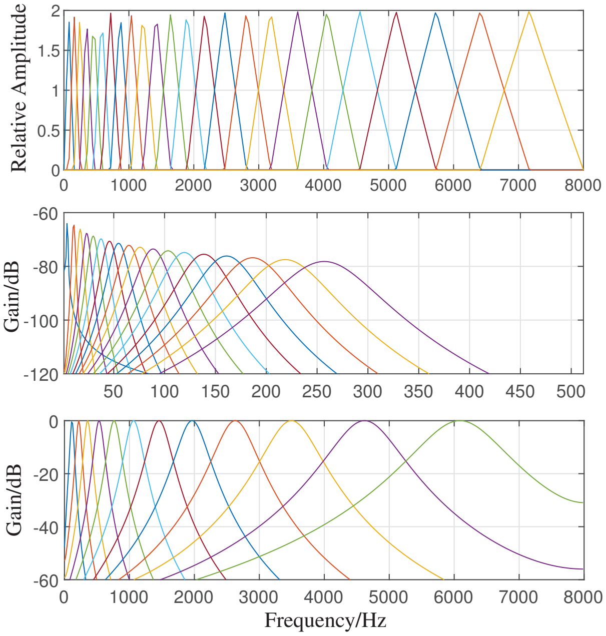

Fig. 1. The frequency response curve of Mel filter bank (top), cochlear filter bank (middle) and Gammatone filter bank (bottom).

In this paper, Melbankm function in MATLAB Voicebox Toolbox was used to calculate Mel filter Banks and the number of Mel filters was set as 24 [26]. The number of cochlear filters was set as 18. And for Gammatone filter, the order of the filter was set as 4 and the number of filters was 64 by the empirical value. Taking signal sampling frequency as 16kHz, frame size as 25ms, the frequency response curves of these filter banks obtained by triangular shaped filters are shown in Fig. 1.

D. Log-Spectrum

In this paper, we adopt the signal spectrogram to define the log-spectrum feature parameters used in this paper [26]:

$$S(i) = \frac{1}{M} \sum_{m=1}^{M} \log |X(m,i)| \tag{7}$$

where, i represents a certain frequency band, M is the number of frames contained in an utterance, and $X(m,i)$ denotes the discrete Fourier transform of the signal in the mth frame. In the experiment, we refer to the previous work and analyzed the information within frequency interval of $0-1200$Hz, which corresponding to low frequency component, namely the first 30 mean of log-spectrum (MLS) coefficients [27].

In addition to MLS coefficients, we also extracted some auditory spectrogram coefficients as an extension of spectrum features, such as robust MFCC (MFCC-R) [26], robust CFCC (CFCC-R), and robust GFCC (GFCC-R). It is worth noticing that applying CFCC-R and GFCC-R features for SER under noise conditions is an additional innovation in this paper. When extracting these features, the operation of DCT transformation was not conducted and the parameters were directly obtained after processing each filter bank. In order to satisfy the structural features of human ears and highlight the dynamic changes of speech signals, the first 12 MFCC coefficients, the first 18 CFCC coefficients, the first 31 GFCCs and their zero-order difference coefficients, first-order difference coefficients and second-order difference coefficients were extracted based on speech frames. At last, the mean of each cepstral and spectrum coefficients was calculated for each utterance. All features used in the experiments are presented in TABLE I.

TABLE I
Feature set.

Feature groups	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
No. dimension	37	37	30	54	18	93	64

IV. Noise Robust Experiments

In this section, the robustness of the features mentioned before for SER were evaluated based on the BP neural network classifier. The number of nodes in the input layer and output layer were determined by the input and output sequence of the SER task. The connection weights between neurons in the input layer, hidden layer and output layer were randomly initialized. The bias of neurons in the hidden layer and the output layer were also randomly initialized. The learning rate was set to 0.1. The log sigmoid function was used as the nonlinear activation function and the additional momentum method in MATLAB Toolbox was adopted to update the weights. Data were first normalized to the scale of $[0, 1]$. Then each emotion utterance was enframed and windowed. Each frame contains 400 sample points and the frame shift includes 160 sample points. The triangular shaped window with equal length as the frame length was used. With an eye to extracting real voice utterances and reducing the computational burden of subsequent processing, endpoint detection was also performed through each utterance. After preprocessing, the feature parameters were extracted as it was described in Section III. During training, 75% data from the extracted features used as the training data and the remaining 25% as the test data. The split of training and testing data was randomized. The trained network that achieved the best test results on clean data was also used to evaluate additive noise data.

A. Experiments on EmoDB

As we can see from the results in TABLE II, for clean speech, the average recognition accuracy rate based on GFCC is higher than that on MLS by 6.17%, MFCC by 4.95% and CFCC by 3.22%. In addition, the average recognition accuracy rate with cepstral coefficients and their difference coefficients is respectively higher than that of spectrum coefficients. For example, taking relative error as an index, the accuracy rate of MFCC surpass MFCC-R by 0.14%, CFCC is over CFCC-R by 2.21%, and GFCC is higher than GFCC-R by 1.26%. Cepstral coefficients have some advantages. First, in terms of algorithm, the feature extraction process of cepstral coefficients includes discrete cosine transform (DCT), which has the superiority on rich signal spectral components and concentrated energy.

TABLE II
EMOTION RECOGNITION RESULTS OF DIFFERENT FEATURES ON EMODB. ACCURACY IN [%].

	Anger	Happiness	Neutral	Sadness	Average
MFCC	92.59	71.43	85.71	100.00	**87.43**
MFCC-R	96.30	73.68	84.21	95.00	87.30
MLS	90.32	70.59	94.44	89.47	86.21
CFCC	100.00	73.68	94.74	88.24	**89.16**
CFCC-R	90.00	71.43	87.50	100.00	87.23
GFCC	92.86	83.33	100.00	93.33	**92.38**
GFCC-R	94.29	77.78	92.86	100.00	91.23

TABLE III
EMOTION RECOGNITION RESULTS UNDER GAUSSION WHITE NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	25.40	25.06	24.04	25.00	24.57	22.58	**46.12**
-5dB	26.21	26.56	**54.41**	25.81	29.05	25.37	41.53
0dB	28.23	29.94	**76.15**	31.59	27.07	32.04	27.82
5dB	34.52	30.26	**70.58**	46.70	26.57	48.08	29.15
10dB	46.21	43.46	**72.75**	62.30	54.36	61.68	45.06
15dB	56.25	60.78	78.98	76.82	76.48	**79.16**	61.35
20dB	61.98	72.92	79.76	80.40	85.95	**88.44**	79.38
25dB	64.02	77.99	80.27	83.33	88.65	**92.09**	86.84
30dB	68.45	81.50	80.27	84.29	90.20	**94.03**	89.05
35dB	79.82	83.75	80.47	85.36	90.60	**94.11**	89.48
40dB	83.45	84.34	80.82	86.11	90.72	**94.71**	90.15

TABLE IV
EMOTION RECOGNITION RESULTS UNDER PINK NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	25.00	23.20	**36.69**	29.11	25.40	28.00	24.19
-5dB	26.97	24.78	**44.07**	34.68	25.86	27.98	22.21
0dB	41.93	28.34	**62.98**	35.09	31.50	31.12	28.33
5dB	49.02	39.97	**70.35**	56.30	51.36	40.60	32.10
10dB	51.87	53.58	**72.50**	63.71	67.16	63.11	46.78
15dB	55.94	63.28	76.81	80.64	**84.31**	82.11	60.80
20dB	60.18	69.14	80.46	81.96	87.92	**91.87**	81.33
25dB	67.24	76.00	80.26	84.73	89.65	**92.16**	89.41
30dB	77.84	79.12	80.47	85.63	89.61	**94.58**	91.45
35dB	84.98	82.35	80.47	86.34	89.52	**94.77**	90.42
40dB	87.00	82.18	81.17	87.09	91.10	**95.02**	90.89

TABLE V
EMOTION RECOGNITION RESULTS UNDER FACTORY NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	25.00	22.39	34.73	25.00	25.00	26.59	**50.94**
-5dB	25.39	25.47	34.35	25.00	27.82	33.30	**67.65**
0dB	30.78	43.26	42.54	25.00	43.96	44.96	**79.81**
5dB	48.88	59.93	68.06	30.00	34.15	62.35	**84.30**
10dB	67.45	70.07	69.29	42.50	40.98	77.08	**88.31**
15dB	78.71	75.80	75.93	62.04	59.04	87.99	**90.02**
20dB	83.78	79.12	80.03	80.33	77.58	**91.10**	90.96
25dB	86.44	82.00	79.59	87.95	85.74	**94.42**	90.61
30dB	88.64	82.18	79.15	87.79	88.43	**94.82**	90.26
35dB	90.44	82.22	79.39	90.24	89.77	**94.86**	90.85
40dB	91.19	82.06	80.37	90.15	90.90	**94.75**	90.49

DCT technique can also achieve a good speech enhancement effect with low computational complexity, while the speech enhancement process can improve the SER performance of feature parameters in a sense. Secondly, since speech signal is a short-term stationary signal, most researchers extract emotional features by using frame processing on speech signal previously. However, the features extracted based on a certain frame are local features, which cannot accurately reflect the dynamic characteristics of emotional speech. Therefore, it is often impossible to build a robust emotional recognition system by simply adopting local features. After framing, by extracting the differential parameters of local features at the statement level and fusing the two statement-level features together, the classification performance can be effectively improved. The experimental results based on EmoDB show that the fused static and difference features improves the recognition rate of locally static features by 1.65% for MFCC, 1.25% for CFCC and 2.57% for GFCC, respectively.

B. Experiments on artificial additive data

The robust classification performance of speech features is tested on the artificial additive noise dataset. In the experiments, the classifier was trained with clean signals and then tested with noisy signals, regardless of speaker dependence. The experimental results on noisy data are tabulated in TABLE III, TABLE IV, TABLE V, and TABLE VI.

From TABLE III to TABLE VI, we can see that the average recognition rates of almost all features drop off along the decrease of SNRs. Under lower SNRs, auditory spectrogram-based coefficients have a better performance than cepstral coefficients. For example, under Gaussion noise condition, MLS has the highest accuracy among other features at 10dB, 5dB, 0dB, and -5dB. Under factory noise condition, GFCC-R achieves a higher accuracy than other features below 20dB. The good property of auditory spectrogram-based coefficients also can be noticed under pink noise and vehicle noise conditions. While cepstral features are able to gain a better performance than auditory spectrogram-based features under higher SNRs. Taking GFCC for example, GFCC achieves the best accuracy rate among the whole extracted features above 20dB under four kinds of noise conditions. This phenomenon

TABLE VI
EMOTION RECOGNITION RESULTS UNDER VEHICLE NOISE CONDITION. ACCURACY IN [%].

	MFCC	MFCC-R	MLS	CFCC	CFCC-R	GFCC	GFCC-R
-10dB	28.48	**53.01**	25.00	25.00	25.00	26.59	50.94
-5dB	34.79	**71.56**	28.04	25.00	27.82	33.30	67.65
0dB	50.41	76.66	45.88	25.00	43.96	44.96	**79.81**
5dB	69.68	80.09	64.54	30.00	34.15	62.35	**84.30**
10dB	77.82	81.99	67.11	55.97	40.98	77.08	**88.31**
15dB	83.09	82.18	75.70	62.04	59.04	87.99	**90.02**
20dB	87.37	82.57	78.84	80.33	77.58	**91.10**	90.96
25dB	90.79	82.37	80.14	87.95	85.74	**94.42**	90.61
30dB	91.42	82.76	80.39	87.79	88.43	**94.82**	90.26
35dB	91.11	83.11	79.61	90.24	89.77	**94.86**	90.85
40dB	92.44	83.11	79.65	90.15	90.90	**94.75**	90.49

is in accord with the results on EmoDB, where GFCC has the highest recognition rate over all features and the difference value of recognition accuracy based on different features is not greater than 0.5. Though cepstral features such as MFCC is a sub-band energy-based feature which has good representations of speech spectral information, it is sensitive to noise and thus is less efficient in distinguishing emotions from noisy speech.

V. CONCLUSIONS

It is a novelty for this paper to conducted a comparative study of the auditory spectrogram-based features and cepstral features on SER with two types of speech data sources (clean speech, and noisy speech). Results show that auditory spectrogram-based features have a robuster performance than cepstral features under lower SNRs. While cepstral features are able to gain a better performance than auditory spectrogram based features, in terms of classification accuracy under higher SNRs, which is contrary to the classification results under lower SNRs. In future work, we plan to explore the detailed compensation relationship between auditory spectrogram-based features and cepstral features when coping with emotion recognition tasks, and design a fusion scheme for SER with better noise robustness.

REFERENCES

[1] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres,”, Speech Communication, vol. 49, no. 2, pp. 98-112, 2007.

[2] D. Tomotsune, M. Shirai, Y. Takihara, and K. Shimada, “Detecting deception: the promise and the reality of voice stress analysis,” Journal of Forensic Sciences, vol. 27, no. 2, pp. 340-51, 1982.

[3] K. Forbes-Riley, and D. Litman, “Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor,” Speech Communication, vol. 53, no. 9-10, pp. 1115-1136, 2011.

[4] H. W. Loellmann, H. Barfuss, A. Deleforge, and S. Meier, “Challenges in Acoustic Signal Enhancement for Human-Robot Communication,” Speech Communication, pp. 1-4, 2014.

[5] S. Zhao, Y. Yang, and J. Chen, “Effect of Reverberation in Speech-based Emotion Recognition,” 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE), pp. 1-5, 2018.

[6] J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, “Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition,” Proceedings of the 2016 Acm Multimedia Conference, pp. 670-674, 2016.

[7] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, “AN UNSUPERVISED FRAME SELECTION TECHNIQUE FOR ROBUST EMOTION RECOGNITION IN NOISY SPEECH,” 26th European Signal Processing Conference (EUSIPCO), 2018.

[8] A. K. Alimuradov, A. Y. Tychkov and P. P. Churakov, “A Method for Noise-Robust Speech Signal Processing to Assess Human Psycho-Emotional State,” 3rd School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), pp. 6-8, 2019.

[9] MCA. Korba, H. Bourouba, and R. Djemili, “FEATURE EXTRACTION ALGORITHM USING NEW CEPSTRAL TECHNIQUES FOR ROBUST SPEECH RECOGNITION,” MALAYSIAN JOURNAL OF COMPUTER SCIENCE, vol. 33, num. 2, pp. 90-101, 2020.

[10] U. Kumaran, SR. Radha, SM. Nagarajan, and A. Prathik, “Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN,” INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY, 2021.

[11] M. Russo, M. Stella, S. Marjan, and V. Pekić, “Robust Cochlear-Model-Based Speech Recognition,” Computers, vol. 8, no. 1, 2019.

[12] A. Mansour, and Z. Lachiri, “A comparative study in emotional speaker recognition in noisy environment,” IEEE/ACS 14th International Conference on Computer Systems and Applications, pp. 980-986, 2017.

[13] X. J. Zhao, Y. Shao, and D. L. Wang, “CASA-Based Robust Speaker Identification,” IEEE Transactions on Audio Speech and Language Processing, vol. 20, no. 5, pp. 1608-1616, 2012.

[14] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognition of Nonprototypical Emotions in Reverberated and Noisy Speech by Nonnegative Matrix Factorization,” Eurasip Journal on Advances in Signal Processing, vol. 2011, no. 1, pp. 1-16, 2011.

[15] MX. Hou, JX. Li, and GM. Lu, “A supervised non-negative matrix factorization model for speech emotion recognition,”, SPEECH COMMUNICATION, vol. 124, pp. 13-20, 2020.

[16] SR. Bandela, and TK. Kumar, “Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition,” APPLIED ACOUSTICS, vol. 172, 2021.

[17] V. Mitra, A. Tsiartas, and E. Shriberg, “Noise and reverberation effects on depression detection from speech,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2016.

[18] R. Sun, and II E. Moore, “Investigating the robustness of teager energy cepstrum coefficients for emotion recognition in noisy conditions,” 2012.

[19] P. K. Aher, S. D. Daphal, A. N. Cheeran, “Analysis of Feature Extraction Techniques for Improved Emotion Recognition in Presence of Additive Noise,” 2016.

[20] S. Hamsa, I. Shahin, Y. Iraqi, and N. Werghi, “Emotion Recognition From Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier,” IEEE ACCESS, vol. 8, pp. 96994-97006, 2020.

[21] A. R. Avila, J. Monteiro, O. Douglas, and T. H. Falk, “Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks,” in IEEE International Symposium on Signal Processing and Information Technology, 2017.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A Database of German Emotional Speech,” Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech), pp. 1517-1520, 2005.

[23] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech and Signal Processing, 1980.

[24] S. S. Stevens, “A Scale for the Measurement of the Psychological Magnitude Pitch,” Journal of the Acoustical Society of America, vol. 8, no. 3, pp. 185-190, 1937.

[25] Q. Li and Y. Huang, “An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 6, pp. 1791-1801, 2011.

[26] E. M. Albornoz, D. H. Milone, and H. L. Rufine, “Feature extraction based on bio-inspired model for robust emotion recognition,” Soft Computing, vol. 21, no. 17, pp. 5145-5158, 2017.

[27] E. M. Albornoz, D. H. Milone, and H. L. Rufine, “Spoken emotion recognition using hierarchical classifiers,” Computer Speech and Language, vol. 25, no. 3, pp. 556-570, 2011.