INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A novel frequency warping scale for speech emotion recognition

Premjeet Singh, Goutam Saha

Dept of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, India

premsingh@iitkgp.ac.in, gsaha@ece.iitkgp.ac.in

Abstract

We investigate an optimised non-linear frequency warping scale for speech emotion recognition (SER). The proposed scale maps the speech spectrogram onto another time-frequency domain which is invariant to speaker-specific variations. Generally, the famous mel-scale designed on human audio perception is considered the de facto standard of frequency warping. However, designed mainly for speech recognition, the generalisability of mel on other speech processing tasks is debatable. Our experiments show that an emotion-specific scale designed on an SER database outperforms the standard mel-scale. Along with performance improvement, the proposed approach also provides insight into the emotion-relevant frequency regions for SER. Despite the database-dependent design of our approach, we find that the scale obtained from our experiments also shows SER performance improvement when tested on two other databases.

Index Terms: Speech emotion recognition, Non-linear frequency warping, Constant-Q transform

1. Introduction

Speech emotion recognition (SER) includes a pattern recognition task of predicting emotional state of spoken utterances. With applications in fields like patient monitoring, autonomous driving vehicles, customer care services, etc., a complete speech emotion aware system is expected to bridge the efficiency gap in human-computer interaction systems [1, 2, 3]. However, the lack of complete understanding of emotion-relevant characteristics of speech and their corresponding correlations is a major challenge that prevents SER supremacy [3, 4, 5, 6, 7].

Along such lines, an optimised time-frequency representation (TFR) that emphasises emotion-relevant spectral regions continues to be a debatable aspect of SER [1, 8]. Inspired by the non-linear (logarithmic) processing of frequencies in human cochlea, several methods use different types of logarithmic warping scales (e.g., mel, bark, equivalent rectangular bandwidth (ERB), constant-Q) on the standard short-time Fourier transform (STFT) to obtain a TFR. Application of such a scale provides a frequency compressed representation with reduced spectral variations [9]. These variations originally stem from the differences in vocal tract anatomy, speaking style, cultural background, spoken language, gender, age, emotional state, etc. [10, 11, 12]. The non-linear scale applies a bilinear transformation that maps the STFT onto a frequency-warped space where these differences are normalised, atleast to some extent [12, 13, 14]. Therefore, search for an SER-optimised scale involves a warping function that normalises speaker-dependent variations while preserving the emotional attributes.

In speech processing domain, one of the most successful warping scale is the mel-scale [15]. The features obtained after applying mel-filterbank on STFT are termed mel-frequency spectral coefficients (MFSC). Authors in [16] found that such transformation provides invariance to temporal and spectral deformations leading to a stable TFR. However, the mel non-linearity was originally designed for speech recognition and is not the most optimum scale for every speech-related task, as found by studies in anti-spoofing [17], and SER [18, 19]. Also, a perception-based filterbank does not guarantee a best fit for every speech-related task [20]. Studies show that a data-driven TFR obtained by employing filterbank with denser sampling of higher signal-to-noise ratio (SNR) frequency regions outperform mel-scale [21, 22]. These observations question the notion of a generalised scale for different speech tasks and show the requirement of a speech-task-specific non-linear warping scale.

Recently, the advent of deep neural networks (DNNs) has inspired automatic feature learning and an end-to-end approach to SER. Several studies use STFT or raw speech directly with a DNN model so as to automatically learn an optimised TFR [20, 23, 24]. Although end-to-end DNN-based models have been shown to outperform the traditional handcrafted features, they also have disadvantages, such as the requirement of a large database [5, 25], and lack of physical insight into task-relevant characteristics [26]. The unavailability of large emotion databases [25, 27] and limited understanding of emotion-relevant speech characteristics make these disadvantages significant from SER perspective. According to [28], generalisation across speech variations can be achieved by: selecting the most task-relevant acoustic cues (analogous to handcrafted features), and learning speaker-specific variations (analogous to DNN modeling). Therefore, an approach that combines domain-knowledge-based feature extraction with DNN-based enrichment of the extracted information can help to achieve a best-of-both-worlds solution. This becomes our motivation to design a handcrafted representation for improved emotion-information extraction and feed it to DNN for further generalisation across speaker-related variations.

In this work, we investigate the following questions: is there an emotion-specific frequency warping scale? Can the standard constant-Q/mel scale be modified to obtain such a scale? Would such a scale developed from a database give poor SER performance on another database? To answer these, we propose modifications on the constant-Q scale, i.e., constant-Q transform (CQT), to obtain a novel TFR for SER. Our choice of CQT, to begin with, stems from its higher low frequency resolution as compared to the mel-scale [17]. As pitch and lower pitch harmonics are found to be more emotion relevant [29], higher resolution around lower pitch harmonics makes the constant-Q scale a better contender for SER [30]. We also compare the performance of the proposed approach by applying similar modifi-

10.21437/Interspeech.2023-1600

cations on the mel-scale. The optimised TFRs are then fed to a DNN classifier backend for performance evaluation.

2. SER specific frequency warping

Apart from stability against temporal deformations [16], frequency warping also introduces stability to speaker-specific variations. A logarithmic scale transforms the speaker-specific uniform scaling of frequency components in the spectrogram to frequency translation in the warped non-linear TFR [31]. However, considering the speaker-specific effects to be *uniform* is a rudimentary assumption. Rather a frequency-dependent non-uniform scaling can better represent the vocal tract differences across speakers [31]. In the following subsections, we show how the constant-Q scale converts uniform spectral scaling to spectral translation. We then discuss modifications that deal with non-uniform spectral scaling for an improved TFR.

2.1. Constant-Q transform

The constant-Q transform of a signal $x(n)$ is given as [17],

$$X^{\mathrm{CQT}}[k,n] = \sum_{j\,=\,n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j)a_k^*(j-n+N_k/2) \quad (1)$$

where k is the CQT frequency bin, N_k is the window length, $\lfloor . \rfloor$ denotes the rounding-off to the nearest integer towards negative infinity, and $a_k^*(n)$ denotes the complex conjugate of the basis function for k^{th} CQT bin. The filter centre frequencies in constant-Q transform are given by $f_k = f_{\min} 2^{\frac{k-1}{B}}$ where $f_{\min}$ is the lowest frequency bin, and B is the number of frequency bins per octave. The constant-Q scale employs a binary logarithmic ($\log_2$) non-linearity which leads to denser filter arrangement at low frequencies as compared to the standard mel-scale. To observe the effect of constant-Q scale on speaker-related frequency transposition, we consider two speakers, X & Y, with frequency components related by a uniform scaling factor α. The constant-Q warping applied on linear frequency ω can also be written as $f(k) = F\left(\omega = 2^k\right)$. Then,

$$\begin{aligned} f_X(k) = F_X\left(\omega = 2^k\right) &= F_Y\left(\alpha\omega\right) \\ = F_Y\left(\alpha\, 2^k\right) &= F_Y\left(2^{(k+\log_2\alpha)}\right) \\ &= f_Y(k+\log_2\alpha). \end{aligned} \quad (2)$$

Thus the uniform frequency scaling between speakers is transformed into frequency translation (shift) in the warped domain. The classifier now only needs to learn invariance against these frequency translations, instead of the scaling factor, for proper classification. However, the frequency transposition appearing due to the vocal tract differences across speakers can be better modelled by a non-uniform frequency-dependent scaling [31]. Therefore, the Eq. 2 analogous representation of frequency-dependent scaling ($\alpha(\omega)$) is, $F_X(\omega) = F_Y\left(\alpha(\omega)\omega\right)$. We now require a new frequency-dependent non-linear warping function that can map the $\alpha(\omega)$ in the original domain (spectrogram) to a translation in the warped domain. Mathematically,

$$\begin{aligned} f_X(k) = F_X(\omega = \mu) &= F_Y(\alpha(\omega)\mu) \\ &= f_Y(k+\eta), \end{aligned} \quad (3)$$

where μ is the required new non-linear warping function (scale) and η is the corresponding shift in the warped frequency domain.

2.2. Gaussian modulated constant-Q scale

In CQT, for a fixed value of bins per octave, the separation between filter centre frequencies increases by $2^{\frac{1}{B}}$. This causes dense filter placement at low frequencies (e.g., 4 filters below 80 Hz in our CQT configuration, cf. Section 3.2), where speech information is very less. As the SER literature reports greater emotion relevance of pitch and lower pitch harmonics [29, 30], filter placement below the pitch frequency range is not desirable.

One solution to the above-mentioned shortcoming is the shifting of filter centre frequencies, especially those at low frequencies, towards higher side. This would increase the filter density, and hence, frequency resolution around lower pitch harmonics benefiting SER. To move the centre frequencies, we add a Gaussian function to the constant-Q scale so as to shift the filters without distorting the continuity of the original scale. The new modified non-linear scale is then given as,

$$f_k = f_{\min} 2^{\frac{k-1}{B}} + A e^{-\frac{(k-\delta)^2}{\beta}}, \quad (4)$$

where, A is amplitude, δ is shift, and β is bandwidth of the Gaussian function. We choose a Gaussian function so that a localised deformation of the scale can be obtained with fewer parameters. In Section 3.2, we describe how the values of these parameters are chosen to optimise SER performance.

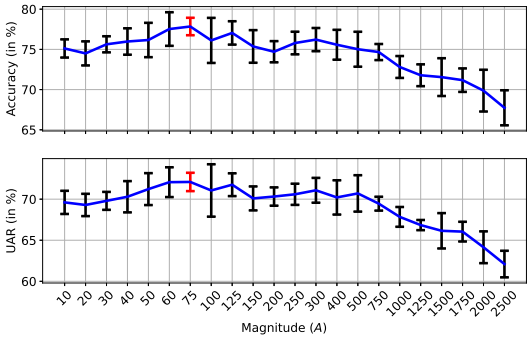
The Eq. 2 analogous expression for the modified scale (ignoring parameters B, A, and β for simplicity) is given as,

$$\begin{aligned} f_X(k) = F_X\left(\omega = 2^k + e^{-(k-\delta)^2}\right) &= F_Y(\alpha\omega) \\ = F_Y\left(\alpha\left(2^k + e^{-(k-\delta)^2}\right)\right). & \end{aligned} \quad (5)$$

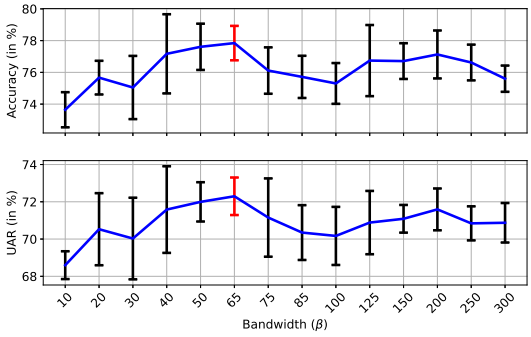
To simplify further, we approximate the effect of Gaussian on the original scale with two exponential functions, i.e.,

$$\begin{aligned} F_Y\left(\alpha\left(2^k + e^{-(k-\delta)^2}\right)\right) &\approx \begin{cases} F_Y\left(\alpha\left(2^k + e^{(k-\delta)}\right)\right), & \text{for } k \le \delta \\ F_Y\left(\alpha\left(2^k + e^{-(k-\delta)}\right)\right), & \text{for } k > \delta \end{cases} \\ &= \begin{cases} F_Y\left(2^{(k+\log_2\alpha)} + e^{(k-\delta+\ln\alpha)}\right), & \text{for } k \le \delta \\ F_Y\left(2^{(k+\log_2\alpha)} + e^{-(k-\delta-\ln\alpha)}\right), & \text{for } k > \delta. \end{cases} \end{aligned} \quad (6)$$

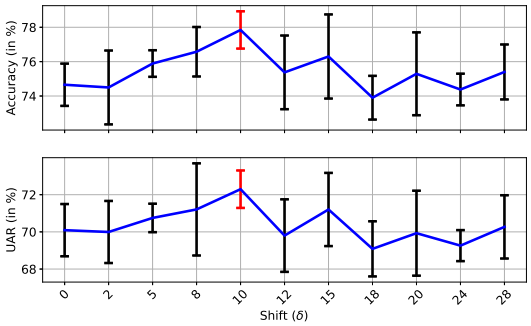
Similar to Eq. 2, the uniform scaling factor α in the original frequency domain is mapped to translation in the warped frequency domain. However, the translation does not take place by a single constant factor only but consists of two separate terms: a constant factor $\log_2\alpha$ and a factor dependent on the warped frequency scale ($\delta - \ln\alpha$ or $\delta + \ln\alpha$). The use of exponentials for piece-wise approximation of the Gaussian represents a comparatively reduced non-linearity (increased slope) in the modified scale before δ and vice-versa. This varying non-linearity in the modified scale leads to a *non-uniform* translation in the warped frequency domain. We relate this to the scale μ in Eq. 3 that transforms an underlying non-uniform frequency-dependent scaling factor $\alpha(\omega)$ to an additive term η. Hence, the Gaussian modified scale results in potentially improved invariance to non-uniform speaker-specific variabilities. In Section 4, we further discuss the SER relevance of the modified scale designed with the optimised values of Gaussian parameters.



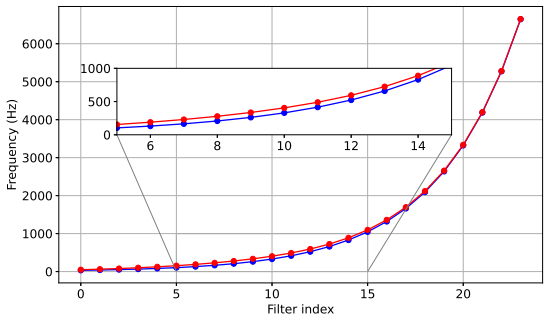
(a) ***Step 1***: *Gaussian magnitude (A) vs performance. Here both* $\beta = 65.0$ *and* $\delta = 10.0$ *are empirically selected.*



(b) ***Step 2***: *Gaussian bandwidth* (β) *vs performance. Here* $A = 75.0$ *is the optimised value from step 1 and* $\delta = 10.0$ *is the empirical value.*



(c) ***Step 3***: *Gaussian shift* (δ) *vs performance. Here both* $A = 75.0$ *and* $\beta = 65.0$ *are optimised values obtained from previous two steps.*



(d) *Comparison between original (blue) and Gaussian modified scale (red).*

Figure 1: *Subfigures (a), (b), and (c) show the optimisation of Gaussian parameters on EmoDB database. Starting from an empirically selected set of values, we first optimise A keeping* β *and* δ *fixed. Next, we optimise* β *by fixing the optimum value of* A *and empirically selected* δ*. Finally,* δ *is optimised using the optimum values of the other two parameters. Experiments are repeated five times and their average and standard deviation is reported. The highlighted part (in red) shows the values with best performance. Subfigure (d) shows the comparison between the original constant-Q and the modified scale.*

Table 1: *Summary of the speech corpora used in the experiments (M:F = Male-Female speakers ratio).*

Databases	# Speakers	# Emotions
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [32]	24 (M:F = 1:1)	8 (Calm, Happy, Sad, Angry, Neutral, Fearful, Surprise, and Disgust)
Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [33]	10 (M:F = 1:1)	4 (Happy, Angry, Sad, and Neutral)
Berlin Emotion Database (EmoDB) [34]	10 (M:F = 1:1)	7 (Anger, Sad, Boredom, Fear, Happy, Disgust, and Neutral)

3. Experimental framework

3.1. Databases used

We use three publicly available SER databases for performance comparison of different non-linear scales. Table 1 provides a brief summary of the databases.

3.2. Parameter values for feature extraction

Following [30], we employ CQT with $f_{\min} = 32.7$ Hz, 3 bins per octave over a total of 8 frequency octaves, and a hop length of 64 samples. To obtain the optimum values of Gaussian parameters A, δ, and β, we first select an empirical value of $A = 10$, $\delta = 10$, and $\beta = 65$. Then, keeping δ and β fixed, we evaluate SER performance for different values of A. In the next step, keeping the optimum A and empirically chosen δ as fixed, we evaluate performance for different values of β. The same procedure is repeated to find δ by replacing the other two parameters with their obtained optimum values. The parameter optimisation is performed on the EmoDB database and the obtained values are then used to evaluate performance on other databases. Regarding selection of initial empirical values, we choose $\delta = 10$ so as to keep the deformation centralised more towards low-frequencies. The choice of $\beta = 65$ ensures greater shifting of low-frequency filters (filters around lower pitch harmonics) towards high frequencies as compared to other values of β. We did not perform a complete grid search to avoid experiments with SER irrelevant parameter combinations, e.g., parameter values that have more effect on high-frequency filters. In fact, our experiments are more focused on investigating whether the scale deformation has some generalised positive SER consequence, and less focused on designing a highly-specialised database-dependent scale. We use *LibROSA* [35] and *nnAudio* [36] libraries to compute MFSC, CQT, and their corresponding modified versions.

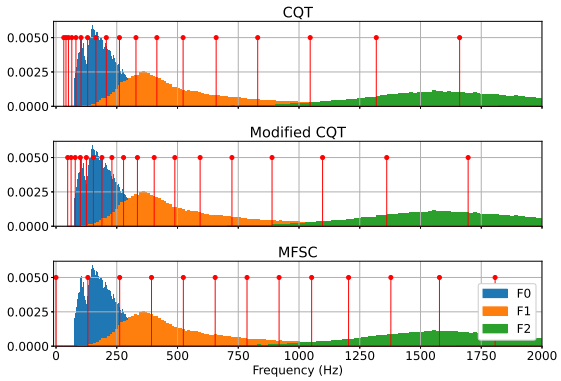


Figure 2: *Comparison of filter density in CQT, modified CQT, and MFSC. The red vertical lines show filter bins for corresponding TFR. The filter bins are plotted over the histogram of pitch (F_0), first formant (F_1), and second formant (F_2) of all utterances in EmoDB.*

Table 2: *Employed DNN architecture. '$\times N$' shows the repeated connection of N layers. FC = Fully-connected.*

Layer	Filters	Kernel Size (Freq. $\times$ Time)	Maxpool (Freq. $\times$ Time)
Conv 2D ($\times 4$)	128	3×5	2×2
Conv 2D ($\times 2$)	128	3×5	-
BiLSTM ($\times 2$)	128	-	-
Attention	-	-	-
Pool			
FC	128	-	-
Softmax	# Classes	-	-

3.3. Classifier description

Table 2 shows the architecture of the employed DNN model. We use ReLU activation, batch normalisation, dropout of 0.3 on only FC layer, and a mini-batch size of 64. A learning rate value of 0.001 is used with scheduler having decay 0.5 and patience of 5 epochs. The lowest learning rate is set to 2.5e$-$4. *PyTorch* python library is used to implement the DNN architecture.

3.4. Evaluation methodology

We use accuracy and unweighted average recall (UAR) as performance metrics. Accuracy is the ratio of the number of samples correctly classified to the total number of samples in the test set. Due to the *unintuitive* nature of accuracy in scenarios with unbalanced emotion classes, we also employ UAR for performance evaluation [37]. We use leave-one-speaker-out (LOSO) cross-validation framework (utterances of one speaker in test, one in validation, and rest in training) to evaluate performance of every database. For training, we segment the utterance into overlapping chunks of 100 consecutive frames, whereas for testing, the utterances are used as such. The DNN model is trained for 50 epochs and the model that provides the highest validation UAR is finally used for testing.

4. Results & discussion

Figure 1 shows the optimisation of Gaussian parameters on EmoDB database. As small sizes of SER databases cause overfitting, we repeat experiments for every combination of Gaussian parameters five times and average the result. During optimisation of A, we observe a peak in average performance around numerical value 75, which drops with a further increase in A. For β optimisation, peak performance is observed at 65, which is also our initial empirical selection. Similarly, for δ, the optimised value is also found equal to the initially selected empirical value. Figure 1d shows the comparison between the original and modified constant-Q scale (with optimised Gaussian). From Figure 1d, we observe that the major deviation from the original scale that causes performance improvement appears below 1 kHz, justifying greater emotion relevance of low frequencies and our empirical choice of δ and β. Figure 2 shows that the modified scale also has greater filter density over pitch histogram as compared to the original scale. This leads to an increased frequency sampling of emotion-relevant frequency regions in the TFR improving SER performance. The modification also leads to reduced non-linearity (increased slope) and hence reduced warping at low frequencies, i.e., before δ and vice-versa (in our experiments, $\delta = 10$, which corresponds to $32.7 \cdot 2^{\frac{10-1}{3}} = 261.6$ Hz). Table 3 shows the performance of different features on different SER databases. Similar to EmoDB, experiments with other databases are also repeated five times and their average is reported. We observe improvement on every database with the modified constant-Q TFR by an almost similar margin, when compared to original constant-Q scale. The similar margin in improvement indicates a similar emotion-relevant effect, i.e., speaker-invariance, provided by the modified scale on every database. Similar modification on mel-scale also shows improvement on EmoDB and RAVDESS databases. Increased speaker invariance due to the modification also helps mel-scale but only on EmoDB and RAVDESS indicating the scale dependency of the modification. We also observe that plain CQT outperforms MFSC on every database by a noticeable margin. For further experimental validation of speaker-invariance, we also perform a Gaussian mixture model (GMM) based speaker identification experiment on EmoDB. We found that original CQT better identifies speakers than modified CQT, justifying the increased speaker invariance in modified CQT.

Table 3: *Performance comparison across databases.*

Features	Metric (in %)	Databases		
		EmoDB	RAVDESS	IEMOCAP
CQT	Acc.	76.10	55.75	52.80
	UAR	70.71	51.94	48.09
Mod. CQT	Acc.	**77.80**	**57.62**	**54.27**
	UAR	**72.17**	**53.42**	**49.17**
MFSC	Acc.	72.14	25.09	46.80
	UAR	67.57	23.34	39.25
Mod. MFSC	Acc.	73.87	34.23	39.99
	UAR	69.70	29.20	34.31

5. Conclusion

We investigate modifications on the constant-Q frequency warping scale for SER. Our experiments show that a Gaussian function based modification of constant-Q scale, especially around pitch frequencies, provides speaker invariance and improvement in SER performance. The performance improvement obtained by modifications optimised on one database is found to generalise across other databases as well. However, similar modifications on the standard mel-scale did not show consistent performance improvement. This shows scale-dependent nature of the employed modification and inspires further analysis of SER-specific frequency warping. In future, a new function can be investigated that provides more control over deformations than the Gaussian. Cross-corpora evaluation can also be investigated for developing a robust scale that generalises well on data from different domains.

6. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] S. R. Krothapalli and S. G. Koolagudi, *Speech emotion recognition: A review*. Springer New York, 2013, pp. 15–34.

[3] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[4] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, pp. 155–177, 2015.

[5] 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.

[6] L. Guo, L. Wang, J. Dang, E. S. Chng, and S. Nakagawa, "Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition," *Speech Communication*, vol. 136, pp. 118–127, 2022.

[7] S. Jing, X. Mao, and L. Chen, "Prominence features: Effective emotional features for speech emotion recognition," *Digital Signal Processing*, vol. 72, pp. 216–231, 2018.

[8] C. Lu, W. Zheng, H. Lian, Y. Zong, C. Tang, S. Li, and Y. Zhao, "Speech emotion recognition via an attentive time–frequency neural network," *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2022.

[9] R. C. Rose, A. Miguel, and A. Keyvani, "Improving robustness in frequency warping-based speaker normalization," *IEEE Signal Processing Letters*, vol. 15, pp. 225–228, 2008.

[10] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved emotion recognition with a novel speaker-independent feature," *IEEE/ASME Transactions on Mechatronics*, vol. 14, no. 3, pp. 317–325, 2009.

[11] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital Signal Processing*, vol. 110, p. 102951, 2021.

[12] S. Umesh, "Studies on inter-speaker variability in speech and its application in automatic speech recognition," *Sadhana*, vol. 36, pp. 853–883, 2011.

[13] E. B. Gouvea and R. M. Stern, "Speaker normalization through formant-based warping of the frequency scale," in *Proc. Eurospeech*, 1997, pp. 1139–1142.

[14] R. Thirumuru, K. Gurugubelli, and A. K. Vuppala, "Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition," *Digital Signal Processing*, vol. 120, p. 103293, 2022.

[15] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[16] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[17] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[18] P. Singh, G. Saha, and M. Sahidullah, "Non-linear frequency warping using constant-Q transformation for speech emotion recognition," in *Proc. International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–6.

[19] P. Singh, M. Sahidullah, and G. Saha, "Modulation spectral features for speech emotion recognition using deep neural networks," *Speech Communication*, vol. 146, pp. 53–69, 2023.

[20] T. N. Sainath, B. Kingsbury, A. Mohamed, and B. Ramabhadran, "Learning filter banks within a deep neural network framework," in *Proc. ASRU*, 2013, pp. 297–302.

[21] S. Sarangi, M. Sahidullah, and G. Saha, "Optimization of data-driven filterbank for automatic speaker verification," *Digital Signal Processing*, vol. 104, p. 102795, 2020.

[22] K. Paliwal, B. Shannon, J. Lyons, and K. Wojcicki, "Speech-signal-based frequency warping," *IEEE Signal Processing Letters*, vol. 16, no. 4, pp. 319–322, 2009.

[23] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. ICASSP*, 2018, pp. 5089–5093.

[24] D. Tang, J. Zeng, and M. Li, "An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals," in *Proc. INTERSPEECH*, 2018, pp. 162–166.

[25] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "LSSED: A large-scale dataset and benchmark for speech emotion recognition," in *Proc. ICASSP*, 2021, pp. 641–645.

[26] Y. Liang, S. Li, C. Yan, M. Li, and C. Jiang, "Explaining the black-box model: A survey of local interpretation methods for deep neural networks," *Neurocomputing*, vol. 419, pp. 168–182, 2021.

[27] S. G. Koolagudi, D. Rastogi, and K. S. Rao, "Identification of language using mel-frequency cepstral coefficients (MFCC)," *Procedia Engineering*, vol. 38, pp. 3391–3398, 2012.

[28] K. Weatherholtz and T. F. Jaeger, "Speech perception and generalization across talkers and accents," in *Oxford Research Encyclopedia of Linguistics*, 2016.

[29] A. Bakhshi, A. Harimi, and S. Chalup, "Cytex: Transforming speech to textured images for speech emotion recognition," *Speech Communication*, vol. 139, pp. 62–75, 2022.

[30] P. Singh, S. Waldekar, M. Sahidullah, and G. Saha, "Analysis of constant-Q filterbank based representations for speech emotion recognition," *Digital Signal Processing*, p. 103712, 2022.

[31] S. Umesh, L. Cohen, and D. Nelson, "Frequency warping and the mel scale," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 104–107, 2002.

[32] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS One*, vol. 13, no. 5, 2018.

[33] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[34] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[35] B. McFee *et al.*, "librosa/librosa: 0.10.0," Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7657336

[36] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "nnAudio: An on-the-fly GPU audio to spectrogram conversion toolbox using 1D convolutional neural networks," *IEEE Access*, vol. 8, pp. 161 981–162 003, 2020.

[37] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Proc. INTERSPEECH*, 2012, pp. 2242–2245.