Non-linear frequency warping using constant-Q transformation for speech emotion recognition

Premjeet Singh, Goutam Saha, Md Sahidullah

Non-linear frequency warping using constant-Q transformation for speech emotion recognition

Premjeet Singh[1], Goutam Saha[2]
Dept of Electronics and ECE
Indian Institute of Technology Kharagpur, Kharagpur, India
premsingh@iitkgp.ac.in[1], gsaha@ece.iitkgp.ac.in[2]

Md Sahidullah[3]
Université de Lorraine
CNRS, Inria, LORIA, F-54000, Nancy, France
[3]md.sahidullah@inria.fr

***Abstract*—In this work, we explore the constant-Q transform (CQT) for speech emotion recognition (SER). The CQT-based time-frequency analysis provides variable spectro-temporal resolution with higher frequency resolution at lower frequencies. Since lower-frequency regions of speech signal contain more emotion-related information than higher-frequency regions, the increased low-frequency resolution of CQT makes it more promising for SER than standard short-time Fourier transform (STFT). We present a comparative analysis of short-term acoustic features based on STFT and CQT for SER with deep neural network (DNN) as a back-end classifier. We optimize different parameters for both features. The CQT-based features outperform the STFT-based spectral features for SER experiments. Further experiments with cross-corpora evaluation demonstrate that the CQT-based systems provide better generalization with out-of-domain training data.**

***Index Terms*—Speech emotion recognition (SER), Constant-Q transform (CQT), Mel frequency analysis, Cross-corpora evaluation.**

I. Introduction

The *speech emotion recognition* (SER) is the task for recognizing emotion from human speech. The potential applications of SER include *human-computer interaction*, *sentiment analysis* and *health-care* [1]–[4]. Humans naturally sense the emotions in speech while machines find it difficult to characterize them [5], [6]. Techniques proposed till date have significantly increased the machine's ability to recognize speech emotions. However, the task is still challenging mainly due to the presence of large *interpersonal and intrapersonal variability* and the *differences in speech quality* used to train and evaluate the system. The goal of this work is to develop an improved SER system by considering emotion-specific acoustic parameters from speech that are assumed to be more robust to unwanted variabilities.

Previous studies in SER research have shown that spectral and prosodic characteristics of speech contain emotion-related information. Spectral features include *lower formants frequencies* (F1 and F2), *speech amplitude and energy*, *zero crossing rate* (ZCR) and spectral parameters, e.g., like *spectral flux* and *spectral roll-off* [7]–[9]. Prosodic features include *pitch*, *pitch harmonics*, *intonation*, and *speaking rate* [8], [9]. The acoustic front-ends are used with *Gaussian mixture model* (GMM) or *support vector machines* (SVMs) as back-end classifiers for SER tasks [10]. Studies with prosody reveal that high arousal emotions, such as *Angry*, *Happy* and *Fear*, have higher average pitch values with abrupt pitch variations whereas low arousal emotions like *Sadness* and *Neutral* have lower pitch values with consistent pitch contours [1], [11]–[14]. The authors in [15] have reported that recognition accuracy of Anger is higher near F2 (1250-1750 Hz) and that of Neutral is higher near F1 (around 200-1000 Hz). Authors in [16] report that center frequencies of F2 and F3 are reduced in depressed individuals. In [17], the authors report that high arousal emotions have higher mean F1 and lower F2 and high (positive) valence emotions have high mean F2. In [18], authors report discrimination between idle and negative emotions using temporal patterns in formants. In [19], the authors have demonstrated that non-linear frequency scales, such as *logarithmic*, mel and *equivalent rectangular bandwidth* (ERB), have considerable impact in SER performance over linear frequency scale.

Recent works with deep learning methods such as *convolutional neural networks* (CNNs) or CNN with *recurrent neural networks* (CNN-RNNs), use spectrogram or raw waveform as input and have shown impressive results [20]–[22]. These data-driven methods automatically learn the emotion-related representation, however, the role of individual speech attributes in the decision making process is not clear due to the lack of *explainability*. On the other hand, the generalization of these methods remains an open problem, especially when the audio-data for train and test are substantially different in terms of language and speech quality [23].

We address this generalization issue by capturing emotion-related information from speech before processing with a neural network back-end. Given the fact that the low and mid frequency regions of speech spectrum contain pitch harmonics and lower formants that are relevant for emotion recognition, we propose to use a more appropriate approach for time-frequency analysis that produces emotion-oriented speech representation in the first place. Even though the processing with *mel frequency warping* introduces non-linearity in some sense, the power spectrum from the speech is essentially computed with a uniform frequency resolution. We propose to use a time-frequency analysis method called *constant-Q transform* (CQT). This transformation offers higher frequency resolution at low-frequency regions and higher time resolution at high-frequency regions. As the pitch harmonics and lower formants reside in the low-frequency regions of speech spectrum, we

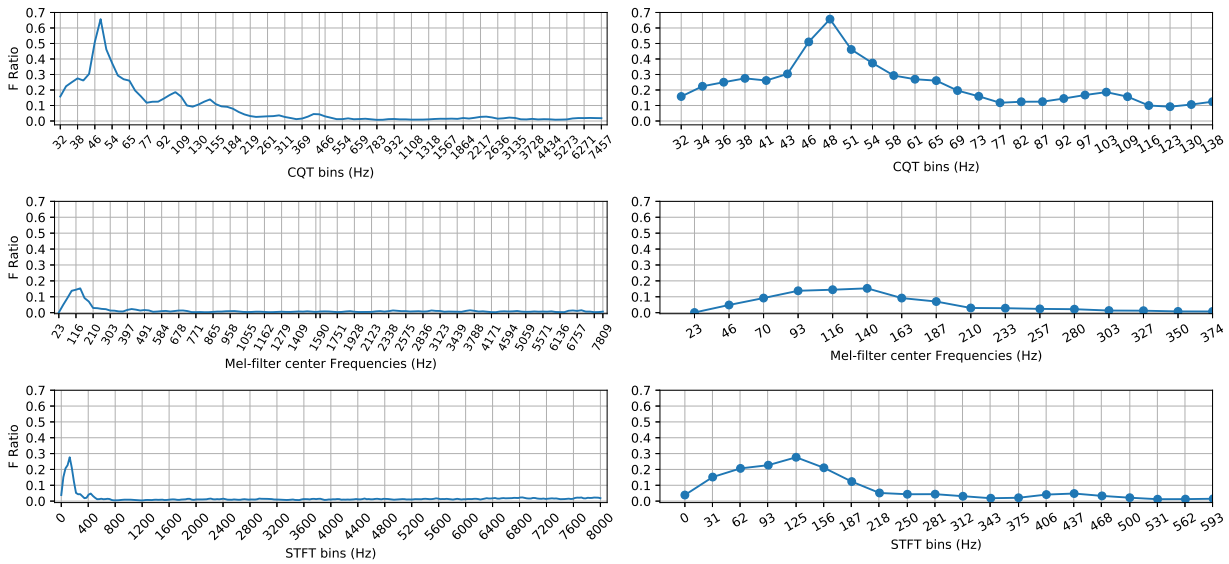

Fig. 1. F-ratios of spectrograms based on CQT (top), mel-filter (middle), and standard STFT (bottom) corresponding to the frequency bins. We use the speech sentences with fixed text 'a02' from EmoDB database (discussed in Section III-A). We select the same text assuming spectra characteristics of emotions to be text-dependent. First column shows the values over the entire frequency range while the second column focuses only on the lower-frequency regions.

hypothesize that keeping high resolution in this region may efficiently capture emotion-related information.

The CQT was initially proposed for music signal processing [24]. Then it was also applied in different speech processing tasks, e.g., *anti-spoofing* [25], [26], *speaker verification* [27] and *acoustic scene classification* [28]. Recently, the CQT has also been studied for SER [29], but without success. This is possibly due to the lack of optimization of CQT parameters and/or the applied end-to-end model fails to exploit the advantages of CQT. Recent studies show that CNN-based models are suitable for SER including cross-corpora evaluation [23], [30]. In this work, we also adopt a CNN-based approach for modeling SER systems. Our main contributions in this work are summarized as follows: (i) We propose a new framework for CQT-based SER by optimizing CQT extraction parameters for reduced redundancy and improved performance, (ii) We investigate CNN architecture known as *time-delay neural networks* (TDNNs) suitable for speech pattern classification tasks [31], [32] for SER, and (iii) We perform cross-corpora evaluation with three different speech corpora to assess the generalization ability of the proposed method. Our results demonstrate that the optimized CQT features not only outperform short-time Fourier transform (STFT) features but also provide better generalization.

II. Methodology

In this section, we discuss the CQT-based feature extraction framework and the TDNN architecture for emotion recognition.

A. Constant-Q transform

The CQT of a time-domain signal $x[n]$ is defined as,

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]e^{-jw_k n}, \qquad (1)$$

where $X[k]$ is the CQT coefficient for k-th frequency bin, $W[k,n]$ is the time-domain window for k-th bin with duration $N[k]$, $x[n]$ denotes the time samples and $w_k = \frac{2\pi Q}{N[k]}$, where Q is the (constant) Q factor of the filter banks [24]. In CQT computation, the window length $N[k]$ varies for different values of k. Hence, $x[n]$ is correlated with sinusoids of different lengths with equal cycles of oscillation. This leads to constant-Q filter bank representation with geometrically spaced center frequencies over frequency octaves. Hence, we obtain a time-frequency representation which has frequency resolution varying from high to low towards increasing frequencies.

The CQT representation of an audio signal depends on the *number of octaves of frequencies* and the *number of frequency bins per each octave*. The number of octaves depends upon the chosen minimum frequency ($F_{\min}$) and the maximum frequency ($F_{\max}$) of operation, and this equals to $\log_2 \frac{F_{\max}}{F_{\min}}$ [25]. The CQT representation with reduced number of total frequency bins over a fixed number of octaves will provide detailed information for lower frequency region with reduced redundancy. Conversely, due to linearly spaced frequency bins, *short-time Fourier transform* (STFT) does not offer this flexibility. Fixing $F_{\min}$ to 32.7 Hz and $F_{\max}$ to *Nyquist frequency* gives us approximately 8 octaves. *Hop length* in CQT computation defines the number of time samples by which CQT computation window moves. The CQT also has resemblance with *continuous wavelet transform* (CWT) which provides variable time-frequency resolution and has been found helpful for SER [33].

During the CQT-based feature extraction process, the CQT coefficients are uniformly resampled and then processed with *discrete cosine transform* (DCT) to compute speech features known as *constant-Q cepstral coefficients* (CQCCs).

We perform *class separability* analysis of the time-frequency representations by computing the F-ratios [34]. The Fig. 1 shows the F-ratio obtained at different frequency bins. The higher F-ratios at lower bins for CQT and STFT show the presence of more discriminative information. The figure also indicates that CQT-spectrogram has more number of discriminative coefficients on an average over others due to higher resolution in low-frequency regions.

B. CNN architecture

The time-frequency representation of speech-like signal is suitable to be used with 1-D CNN, popularly known as TDNN in speech processing literature. Our method is inspired by the TDNN-based *x-vector* system [32] developed for speaker verification task. This processes speech information at *frame* and *segment* level. In frame level, the TDNN captures *contextual information* by applying *kernel* over adjacent frames and by processing each speech frame in an identical manner. This also applies *dilation* in the temporal domain to reduce redundancy and to make it computationally efficient. The frame-level information is processed with several TDNN layers having different kernel sizes and dilation parameters. Finally, *temporal pooling* aggregates frame-level information into segment-level and this is followed by processing with *fully connected* (FC) and *softmax* layer for classification objective. The standard x-vector system computes the segment-level intermediate representation referred as *embeddings* which are further processed with another system for classification. In contrast, our proposed method trains the network in an end-to-end fashion for which the emotion for a test speech is obtained from the output of the trained network.

We empirically optimize the parameters for TDNN architecture. Finally, we use four TDNN layers, followed by statistics pooling with mean and standard deviation, and one FC layer before *softmax*. Table I describes the parameters for different layers.

TABLE I
THE PARAMETERS OF CNN ARCHITECTURE FOR SER.

Layer	Size	Kernel Size	Dilation
TDNN	32	5	1
TDNN	32	3	2
TDNN	32	3	3
TDNN	64	1	1
Statistics Pooling (Mean and SD)	128	-	-
Fully Connected	64	-	-
Softmax	#Classes	-	-

III. EXPERIMENTAL SETUP

A. Speech corpora

In our experiments, we use three different speech corpora which are described in Table II. We downsample speech files at sampling rate of 16 kHz when required. The EmoDB is a German language corpora while RAVDESS and IEMOCAP are in English. For IEMOCAP database, we select only four emotions (Angry, Happy, Sad and Neutral) as some of the emotion class have inadequate data for training neural network models [30]. We perform cross-corpora SER experiments by selecting the same four emotions.

TABLE II
SUMMARY OF THE SPEECH CORPORA USED IN THE EXPERIMENTS. (F=FEMALE, M=MALE)

Databases	Speakers	Emotions
Berlin Emotion Database (EmoDB) [35]	10 (5 F, 5 M)	7 (Anger, Sad, Boredom, Fear, Happy, Disgust and Neutral)
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [36]	24 (12 F, 12 M)	8 (Calm, Happy, Sad, Angry, Neutral, Fearful, Surprise, and Disgust)
Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [37]	10 (5 F, 5 M)	4 (Happy, Angry, Sad and Neutral)

B. Experimental details & evaluation methodology

First, we optimize the parameters of the features on EmoDB. We perform experiments on this corpus using *leave-one-speaker-out* (LOSO) cross validation by keeping one speaker in test. Out of the remaining speakers, we use two of them in validation and seven in training. We also apply five-fold data augmentation by corrupting training set with additive noises and room reverberation effect following the *Kaldi* recipe[1] for x-vector training [32].

We extract features from each speech utterance and discard the non-speech frames with a simple energy-based *speech activity detector* (SAD). We apply utterance-level *cepstral mean variance normalization* (CMVN) before creating the training and validation samples with chunks of 100 consecutive frames. We consider multiple non-overlapping chunks from the speech utterances depending on the length. We use LibROSA[2] python library for feature extraction.

We do not apply chunking for testing and consider the full utterance for computing the test accuracy. We report the final performances with accuracy as well as *unweighted average recall* (UAR). The accuracy is computed as the ratio between the number of correctly classified sentences to the total number of sentences in test. The UAR is given as [38],

$$\text{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}} \tag{2}$$

where A refers to the contingency matrix, A_{ij} corresponds to number of samples in class i classified into class j and K is the total number of classes. As accuracy is considered

[1]https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2
[2]https://librosa.github.io/

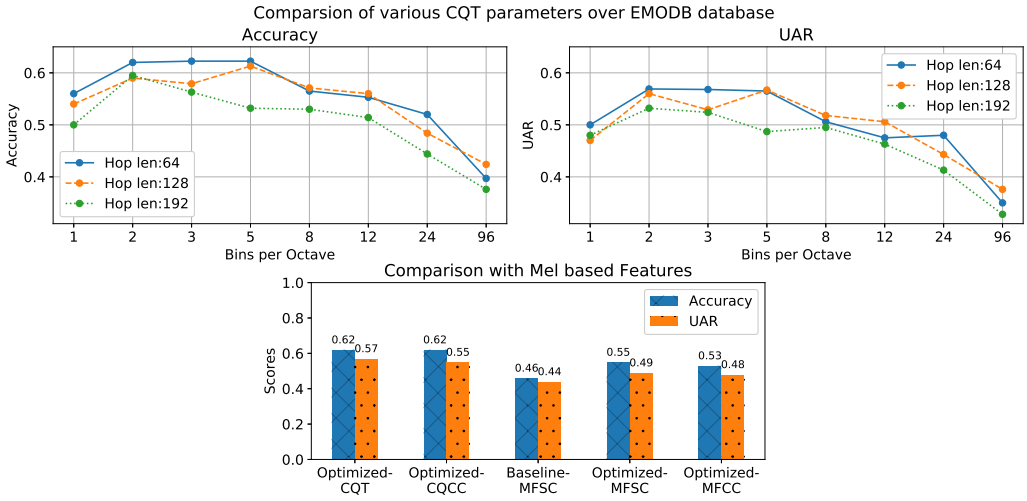

Fig. 2. Performance comparison of CQT for different parameter values. Optimized CQT shows the response of CQT with 3 bins per octave and hop length of 64 samples. Baseline MFSC corresponds to MFSC extraction with standard value of 128 mel-filters and 160 samples hop length, whereas, optimized MFSC has 24 mel-filters with 64 hop. Optimized CQCC and MFCC shown are obtained after applying DCT over optimized MFSC and CQT. The results shown in figure are obtained over EmoDB database only.

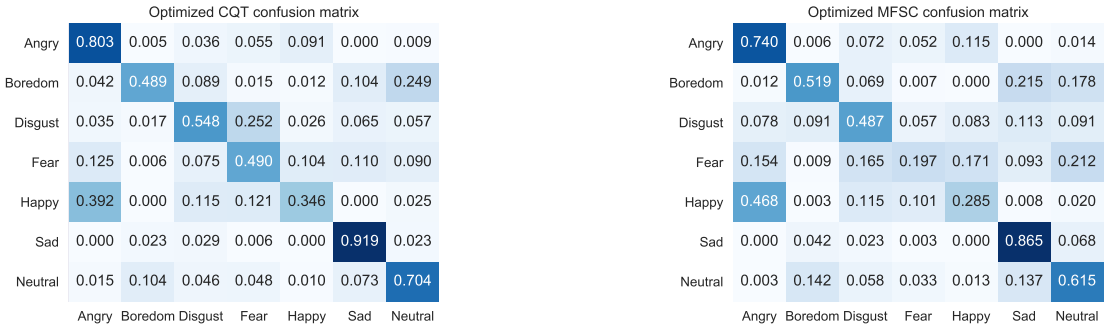

Fig. 3. Confusion matrices for emotion classification experiment with optimized CQT and MFSC features in EmoDB corpus. Given values are the ratio of utterances identified in column class to the total number of utterances in every corresponding row class.

unintuitive for databases with uneven samples across different classes, we optimize the feature extraction parameters based on the UAR metric.

In DNN, we use ReLU activation function and batch normalization for all the hidden layers. For regularization, we apply dropout with probability 0.3 on the FC layer only. We use Adam optimizer with learning rate 0.001. The mini-batch size is 64. We train the models for 50 epochs and finally testing is done with the model which achieves the highest UAR on the validation set. We repeat each experiment multiple times and report the average performance.

IV. Results

A. Experiments on EmoDB

First, we conduct experiments on EmoDB and optimize the CQT parameters. We vary the number of bins per octave from 1 to 96. We also perform the experiments with three different hop lengths: 64, 128, and 192. The top row of Fig. 2 shows the standard accuracy and UAR for CQT. We observe improved performance for lower bins per octave and lower hop length. The performance remains very similar for bins per octaves between 2 and 5. We select 3 bins per octave as the optimum observing the consistency in different runs of the experiment. We fix the hop size 64 as the optimum since the performance is consistently better with this hop size, especially, for lower bins per octave. Since the optimized CQT features use 24 filters and hop length as 64, we apply similar configuration for STFT-based *mel frequency cepstral coefficients* (MFCCs) as well as *mel frequency spectral coefficients* (MFSCs) (i.e., MFCCs without DCT). The SER performances with CQT and STFT-based features are illustrated as a bar plot in Fig. 2. We observe that CQT coefficients as well as CQCCs consistently outperform MFCCs and MFSCs. We also notice that the optimized MFSC outperforms baseline MFSC. The DCT slightly degrades performance in both CQT and STFT-based approaches. We chose the best configuration for both features for the remaining experiments.

Figure 3 shows the confusion matrices obtained for CQT and MFSC in experiments with EmoDB. We observe that CQT is better capable of discriminating emotions such as Fear,

TABLE III
CROSS-CORPORA RESULTS SHOWN IN ACCURACY / UAR. THIS USES OPTIMIZED CONFIGURATION OF MFSC AND CQT. TO HAVE EQUAL NUMBER OF CLASSES, ONLY FOUR EMOTIONS (HAPPY, ANGRY, SAD AND NEUTRAL) ARE CONSIDERED FROM EVERY DATABASE HERE. ALL OTHER PARAMETER SETTINGS REMAIN SIMILAR TO OTHER EXPERIMENTS.

Train on	Test on	MFSC	CQT
EmoDB	RAVDESS	0.41 / 0.44	0.44 / 0.46
	IEMOCAP	0.36 / 0.37	0.38 / 0.39
RAVDESS	EmoDB	0.45 / 0.42	0.48 / 0.48
	IEMOCAP	0.30 / 0.32	0.32 / 0.34
IEMOCAP	EmoDB	0.64 / 0.50	0.63 / 0.50
	RAVDESS	0.38 / 0.39	0.38 / 0.39

Disgust, Sad, Anger and Neutral as compared to MFSC. The CQT-based system yields improved accuracy for Sad, Neutral and Disgust because those emotions are more prominent in low-frequency regions. Performance of Boredom is slightly degraded. Among all the seven emotions, Fear shows the highest gain in performance over MFSC. Happy shows the lowest classification accuracy and a high confusion with Angry.

B. Cross-corpora evaluation

Table III shows the performance obtained after cross corpus testing. The optimized CQT shows better performance than optimized MFSC for most cases except when the train-test pair are IEMOCAP-EmoDB and IEMOCAP-RAVDESS. The obtained results consolidate our hypothesis that CQT helps in better capturing of emotion-dependent information leading to better generality across databases.

V. DISCUSSION AND CONCLUSION

We notice that increasing the frequency resolution at lower frequency regions led to substantial improvement in SER performance. This also confirms that low-frequency region containing pitch harmonics and lower formants convey important emotion-specific information. At the same time, the CQT with lower high-frequency resolution does not degrade the overall SER performance which indicates that high-frequency regions are less important from emotion perspective. Also, better performance with fewer frequency bins in both CQT and MFSC indicates less redundant time-frequency representation is more effective for emotion discrimination. Though STFT with optimized parameters generates spectrograms with higher frequency resolution, the performance degrades most likely due to increased redundancy caused by capturing details of high-frequency region. Cross-corpora evaluation suggests that CQT-based time-frequency representation provides better generalization for SER task with different speech corpora in training and test.

We conclude that CQT is a better choice of time-frequency representation in terms of both recognition performance and generalization ability. However, the SER performance is still poor for real-world deployment. We also gain no improvement over MFSC for all the seven emotions included in EmoDB corpus. This indicates that the time-frequency representation needs further investigation for SER. This work can also be extended by exploring CQT representation with recurrent architecture and attention mechanisms which are lacking within our TDNN framework but found useful for SER.

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, 2020.

[2] Sreenivasa Rao Krothapalli and Shashidhar G Koolagudi, "Speech emotion recognition: a review," in *Emotion recognition using speech features*, pp. 15–34. Springer, 2013.

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

[4] Jesús B Alonso, Josué Cabrera, Manuel Medina, and Carlos M Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9554–9564, 2015.

[5] Rosalind W Picard, *Affective computing*, MIT press, 2000.

[6] Rosalind W Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[7] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Processing*, vol. 22, no. 6, pp. 1154–1160, 2012.

[8] Florian Eyben, Anton Batliner, and Bjoern Schuller, "Towards a standard set of acoustic features for the processing of emotion in speech.," in *Proceedings of Meetings on Acoustics 159ASA*. Acoustical Society of America, 2010, vol. 9, p. 060006.

[9] Florian Eyben 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.

[10] M. Kockmann, L. Burget, and J.H. Černocký, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition," *Speech Communication*, vol. 53, no. 9-10, pp. 1172–1185, 2011.

[11] Carl E Williams and Kenneth N Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[12] Rainer Banse and Klaus R Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614, 1996.

[13] Roddy Cowie and Ellen Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Proc. ICSLP*. IEEE, 1996, vol. 3, pp. 1989–1992.

[14] Suman Deb and Samarendra Dandapat, "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 802–815, 2018.

[15] Sahar E Bou-Ghazale and John HL Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[16] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[17] Martijn Goudbeek, Jean Philippe Goldman, and Klaus R Scherer, "Emotion dimensions and formant position," in *Proc. INTERPSEECH*, 2009, pp. 1575–1578.

[18] Elif Bozkurt, Engin Erzin, Cigdem Eroglu Erdem, and A Tanju Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Communication*, vol. 53, no. 9-10, pp. 1186–1197, 2011.

[19] Margaret Lech, Melissa Stolar, Robert Bolia, and Michael Skinner, "Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, pp. 363–371, 2018.

[20] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[21] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[22] Che-Wei Huang and Shrikanth Narayanan, "Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition," *arXiv preprint arXiv:1706.02901*, 2017.

[23] Jack Parry, Dimitri Palaz, Georgia Clarke, Pauline Lecomte, Rebecca Mead, Michael Berger, and Gregor Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," *Proc. INTERSPEECH*, pp. 1656–1660, 2019.

[24] Judith C Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

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

[26] Monisankha Pal, Dipjyoti Paul, and Goutam Saha, "Synthetic speech detection using fundamental frequency variation and spectral features," *Computer Speech & Language*, vol. 48, pp. 31–50, 2018.

[27] Héctor Delgado et al., "Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification," in *Proc. IEEE SLT*. IEEE, 2016, pp. 179–185.

[28] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification," in *Proc. the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016, vol. 90, pp. 1032–1048.

[29] Dengke Tang, Junlin Zeng, and Ming Li, "An end-to-end deep learning framework for speech emotion recognition of atypical individuals," in *Proc. INTERSPEECH*, 2018, pp. 162–166.

[30] Dias Issa, M Fatih Demirci, and Adnan Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, pp. 101894, 2020.

[31] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[32] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[33] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *Proc. ICSPCS*, 2016, pp. 1–8.

[34] Simon Nicholson, Ben Milner, and Stephen Cox, "Evaluating feature set performance using the F-ratio and J-measures," in *Proc. EUROSPEECH*, 1997, pp. 413–416.

[35] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[36] Steven R Livingstone and Frank 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.

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

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