Analysis of constant-Q filterbank based representations for speech emotion recognition

Premjeet Singh[a,*], Shefali Waldekar[c], Md Sahidullah[b], Goutam Saha[a]

[a]*Department of Electronics & Electrical Communication Engineering*
Indian Institute of Technology Kharagpur, India
[b]*Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France*
[c] *Department of Electrical, Electronics & Communication Engineering, GITAM School of Technology,*
GITAM (Deemed to be University), Bengaluru, India.

Abstract

This work analyzes the constant-Q filterbank-based time-frequency representations for speech emotion recognition (SER). Constant-Q filterbank provides non-linear spectro-temporal representation with higher frequency resolution at low frequencies. Our investigation reveals how the increased low-frequency resolution benefits SER. The time-domain comparative analysis between short-term mel-frequency spectral coefficients (MFSCs) and constant-Q filterbank-based features, namely constant-Q transform (CQT) and continuous wavelet transform (CWT), reveals that constant-Q representations provide higher time-invariance at low-frequencies. This provides increased robustness against emotion irrelevant temporal variations in pitch, especially for low-arousal emotions. The corresponding frequency-domain analysis over different emotion classes shows better resolution of pitch harmonics in constant-Q-based time-frequency representations than MFSC. These advantages of constant-Q representations are further consolidated by SER performance in the extensive evaluation of features over four publicly available databases with six advanced deep neural network architectures as the back-end classifiers. Our inferences in this study hint toward the suitability and potentiality of constant-Q features for SER.

Keywords: Constant-Q filterbank, Constant-Q transform (CQT), Continuous wavelet transform (CWT), Time invariance, Speech emotion recognition (SER).

1. Introduction

Speech emotion recognition (SER) is a machine's ability to recognize emotions in a speech sample. With the evolution of *smart* devices, emotion recognition has become an essential facet of artificial intelligence. To master emotion recognition is a long-sought aim of researchers. The ability of machines to automatically gather human sentiment will

[*] Corresponding author

Email addresses: premsingh@iitkgp.ac.in (Premjeet Singh), swaldeka@gitam.edu (Shefali Waldekar), md.sahidullah@inria.fr (Md Sahidullah), gsaha@ece.iitkgp.ac.in (Goutam Saha)

lead to efficient man-machine interaction. Various applications of SER include assessing a driver's behavior in autonomous driving vehicles, patient monitoring in health-care services, consumer satisfaction in call centers, product analysis [1–3], etc. Although a considerable amount of progress has been made so far, emotion recognition continues to be one of the most challenging domains in speech signal processing [4, 5]. This is mainly due to the differences in emotion expression by different individuals. Although few studies suggest that there are no significant differences between how emotions such as *Happy, Angry, Fearful* and *Sad* are expressed by individuals, there are some variations in the intensity of expression among people from different cultures and background [6–8].

A general SER system contains an emotion-relevant information (feature) extraction module followed by an emotion-class classifier. In terms of emotion information extraction, the unavailability of a standard feature that promises decent emotion information extraction adds to the challenge around SER. Two types of speech features are prominently used in SER, namely *prosodic* and *spectral* features [9, 10]. Prosodic features mainly include pitch, pitch harmonics, intonation, energy, and speaking rate. Spectral features [9–11] include vocal tract resonant frequencies (formants), spectral flux, spectral roll-off, mel-frequency based analysis, etc. Various works which propose specific features for SER generally incorporate techniques to obtain detailed spectral, prosodic, or both spectral-prosodic information [12]. The aim here is to extract distinguishing emotional cues. This information is then fed to a classifier that classifies the speech utterances among different emotion classes.

Another popular approach in SER is combining information obtained from both spectral and prosodic speech characteristics [9]. These methods include extraction of an exhaustive feature set and its statistics over speech segments and whole utterances. They are also termed brute-force methods [13]. Although these show promising results, *curse of dimensionality* becomes an important issue while handling large number of features. Due to the impressive ability of deep learning methods to find the minima of loss functions, many studies employ *deep neural networks* (DNNs) to self-learn optimum features to either develop an end-to-end SER system [14, 15] or to use the learned features for SER after some post-processing [16, 17].

Deep learning networks provide mathematical models that self-learn an end-to-end solution to a particular pattern recognition problem. This makes deep networks very appealing. However, due to the high complexity, deep networks are onerous to interpret [18], hence making it difficult to obtain an insight into the emotionally relevant characteristics of input speech. Further, due to random initialization of parameters, DNNs converge to different local minima even with identical network architecture and optimization criteria. This raises the question of whether the conclusion about interpretability is consistent for all possible solutions.

Deep learning networks contain many parameters and for their proper training massive amount of data is required [19]. As most of the available SER databases are small, transfer learning is frequently used as a substitute approach. However, since the pretrained models are trained on a completely different database, transfer learning limits the exploitation of the full potential of deep networks. Studies also show that deep networks fail to generalize well in out-of-domain SER scenarios [20]. Despite all such disadvantages, DNNs perform better in SER by automatically learning features from speech representations as compared to the traditional handcrafted features applied to other machine learning techniques, such as support vector machine (SVM), linear discriminant analysis (LDA), k-nearest neighbors (k-NN) [16, 17], etc. Therefore, we employ a combination of handcrafted features and deep networks for improved emotion information extraction and enhanced SER performance.

In this work, we use constant-Q filterbank based time-frequency representations with various deep neural network architectures for SER. The time-frequency representation provides the handcrafted descriptor over which the deep neural network further extracts the emotion-relevant information. We analyze the relevance of constant-Q representations from emotion perspective in both time and frequency domains. We then compare constant-Q and mel-scale representations to investigate the effect of different non-linearities on emotion prediction. We also examine and compare the stability of the two features toward time deformations. Additionally, our study shows a striking similarity between the constant-Q feature representations used in this work and first-layer scattering transform coefficients.

In the next section, we review the relevant literature and present our contributions. Section 3 elaborates the features used in our experiments. In Section 4, we analyze the differences between time-frequency representation of features in both time and frequency domains. Section 5.1 describes different classifiers employed in this work. The SER databases used in this study are detailed in Section 5.2. Section 5.3 outlines the experimentation methods. The results and corresponding discussion are given in Section 6, followed by conclusive statements in Section 7.

2. Related works and motivation

2.1. Literature survey

Before the introduction of deep learning into signal processing, handcrafted features were the default choice of the SER researchers. One of the first seminal works on SER used 17 prosodic features, including speaking rate, voiced region duration, and statistics of pitch [21]. The work in [22] used 32 features post feature selection on various statistics of prosodic features, e.g., mean, standard deviation, skewness, kurtosis, etc. Similarly, for discrimination between stressed and normal speech, [23] proposed using spectral features with their autocorrelation-based variants and showed significant improvement. Again, considering only spectral features, [24] used log frequency power coefficients for SER, and they were shown better than mel-frequency cepstral coefficients (MFCCs). In [25], statistics of MFCC on three different phoneme classes of speech signal reportedly improved the performance with increased speech segment length. Experiments performed in [26] showed that modulation spectral features, obtained by applying a separate modulation filterbank on the response of the auditory filterbank, are better in characterising different speech emotions than both MFCC and perceptual linear prediction. In [27], a discrete Fourier-parameters based model was made for SER. Authors observed that frequency harmonics extracted using Fourier analysis, and their first and second-order derivatives, contain adequate information to discriminate among different emotion classes.

Recently, automatic feature extraction using deep neural networks has gained huge interest because of their ability to learn emotion relevant information from speech signals. Those learned features yield competitive SER performance compared to the traditional handcrafted features. End-to-end SER models were proposed with raw emotional speech and *convolutional neural networks* (CNNs) with *long-short term memory*

(LSTM) networks in [14, 28, 29]. Whereas [17] attempted to learn detailed emotion-related information by providing log mel-spectrogram to the input of CNN and then applying *discriminant temporal pyramid matching*. Similarly, [30] used spectrogram of raw speech and glottal waveform as input to stacked denoising autoencoder with bidirectional LSTM (BLSTM) as the classifier. Regarding the suitability of deep learning network architectures for SER, [31] studied 2D CNN, LSTM, and fully-connected (FC) network architectures and reported that 2D CNN network fairs better for SER on IEMOCAP database [32]. CNN can be considered a static classifier that jointly processes many speech frames taken at a time. This led to the inference that SER is more dependent on static or utterance-level speech characteristics than dynamic or frame-level information, which is better processed by LSTM networks [31]. Authors in [20] evaluated the generalization capability of various deep learning architectures in a cross-corpus SER scenario. The study concluded that convolution-based architectures are better for 'in the wild' test conditions.

Among different types of handcrafted features used in SER, spectral features are considered to contain a substantial amount of emotionally-rich information. According to [25], spectral features "convey information on both *what* is being said, and *how* it is being said". A time-frequency representation provides spectral features at different time intervals. In SER, spectrogram, mel-frequency spectral coefficients (MFSCs)[1], and MFCC have been the common choice of time-frequency representation. Spectrogram and its mel-filter variants represent energy distribution across both time and frequency dimensions [33]. Such representations provide details about the temporal spread of emotional patterns in the speech signal. However, these time-frequency representations were less effective without post-processing using phoneme class-wise spectral feature extraction in [25].

Regarding the frequency domain localization of emotion information in time-frequency representation, several studies have shown that the low-frequency regions of the speech spectrum contain important emotion-related details compared to the high-frequency region [34–37]. These studies reveal the relevance of prosody features for SER and report that the emotions with higher arousal, such as *Angry, Happy,* and *Fear*, have higher average pitch frequency. In contrast, emotions with lower arousal, for example, *Sad*, and *Neutral* have lower pitch values and consistent pitch contours.

Authors in [23] reported that *Neutral* has better recognition rate around the first formant frequency F1 (200-1000 Hz) while around the second formant frequency F2 (1250-1270 Hz), the recognition accuracy of *Anger* is higher. The authors in [38] found the center frequencies of formants F2 and F3 to reduce in depressed individuals. In [39], it was shown that emotions with higher arousal have higher values of mean F1 and lower values of F2, whereas positive valence emotions have higher mean F2. Some discrimination between idle and negative emotions was shown using the temporal patterns of first two formant frequencies in [40]. Authors in [41] demonstrated that non-linear frequency scales, such as logarithmic, mel, and equidistant rectangular bandwidth (ERB), positively impact SER performance when compared with linear scale. Hence, the studies performed in SER literature hint at the potential emotion salience of low-frequency speech regions.

[1] We refer to mel-frequency spectral coefficients as MFSC in this work. It is also represented as mel-spectrogram in some works. Another equivalent representation is mel-frequency log energy or MFLE.

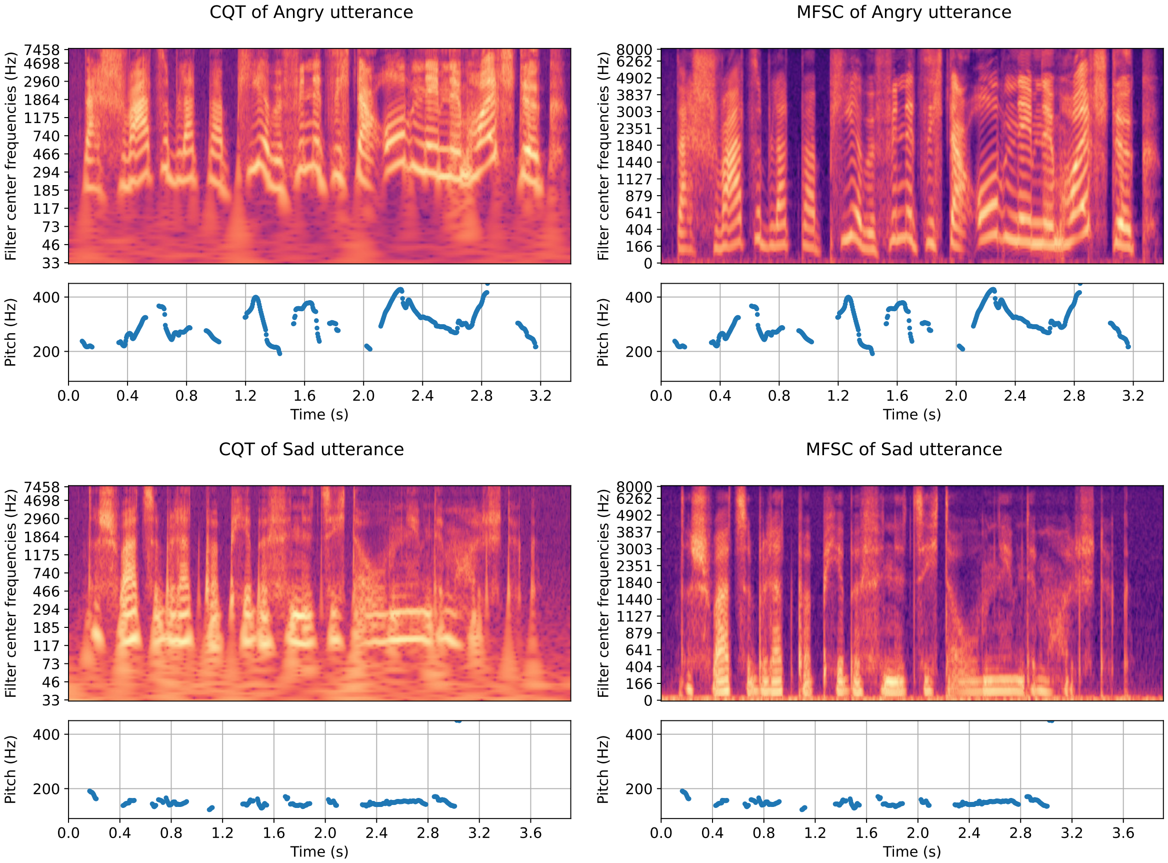


Fig. 1: MFSC and CQT plots of *Angry* and *Sad* utterances of Sentence a05 by Speaker 09 from EmoDB database. CQT representation provides higher number of frequency bins at lower frequencies, hence increasing the low-frequency resolution. This leads to improved pitch resolution in CQT as compared to MFSC, especially in *Sad* emotion where pitch values are low. Another observation in CQT is the smearing of low-frequency components in time, increasing time-invariance.

2.2. Motivation and contributions

The existing SER literature draws attention to the importance of low frequencies of speech signals for emotion recognition. To exploit this characteristic, a non-linear time-frequency representation is required that can emphasize the low-frequency regions. The mel-scale warping in well-known mel-frequency based analysis introduces a logarithm-based non-linearity which provides some emphasis on lower frequencies. To further improve the resolution, we propose the use of constant-Q filterbank based non-linear scale for SER. The constant-Q transform (CQT) applies constant-Q filters to offer higher frequency resolution in low-frequency regions and higher time resolution in high-frequency regions. We hypothesize that, since the pitch harmonics and lower formants, which play a major role in emotion discrimination, reside in the speech spectrum's low-frequency regions, having a higher resolution in this region would capture emotion-related information more efficiently. Authors in [42] also argue that in CQT, the pitch information is visible and harmonics are well separated. Figure 1 compares CQT and MFSC time-frequency representations for two different emotions from EmoDB database. CQT was originally proposed for music processing [43], after which it was successfully applied to other audio processing applications, like anti-spoofing [44, 45], speaker verification, [46] and acoustic scene classification [47, 48]. In [29], CQT was also studied for SER, but no significant improvement was observed. The reason could be the end-to-end model's inability to exploit CQT completely or the inappropriate choice of CQT computation parameters in the experiments.

Another transform that provides constant-Q filter based structure is the continuous wavelet transform (CWT) [49]. CWT also gives varying frequency resolution, similar to CQT, by utilising different scale values of the wavelet basis function. Hence, we also use the CWT time-frequency representation and compare it with MFSC representation for SER. The difference in CQT and CWT then lies only in how time-frequency representation is computed. This difference also helps us in analysing the importance of time-invariance in feature representation for SER. Although CQT is very similar to CWT, the former provides non-redundant features and is, therefore, better suited for a varying resolution time-frequency representation. Also, the CWT has been found superior than mel-filter based techniques for SER in different works, e.g., [50–52], and [53].

In this work, we extend our preliminary study performed in [54] with an in-depth analysis focused on the advantages of constant-Q representations for SER and its comparison with the mel-scale features. We also extensively evaluate features on different databases, using six different DNN architectures. The following are the main contributions of this paper:

- Detailed time-domain analysis of CQT time-frequency representation for SER and comparison with MFSC.
- Time-domain analysis of CWT for SER and its comparison with CQT and MFSC.
- Frequency-domain investigation of differences between mel and constant-Q filter-bank representation.
- Comparison of the performances obtained with different deep neural network architectures for SER.

3. Description of time-frequency representations

3.1. Constant-Q transform

The constant-Q transform of a time-domain signal $x[n]$ is given as [55],

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

where, k denotes the CQT frequency index, $\lfloor . \rfloor$ denotes the rounding-off to nearest integer towards negative infinity and $a_k^*(n)$ is the complex conjugate of the CQT basis function for k^{th} CQT bin. The CQT basis, or the time-frequency *atom*, is a complex time-domain waveform given as,

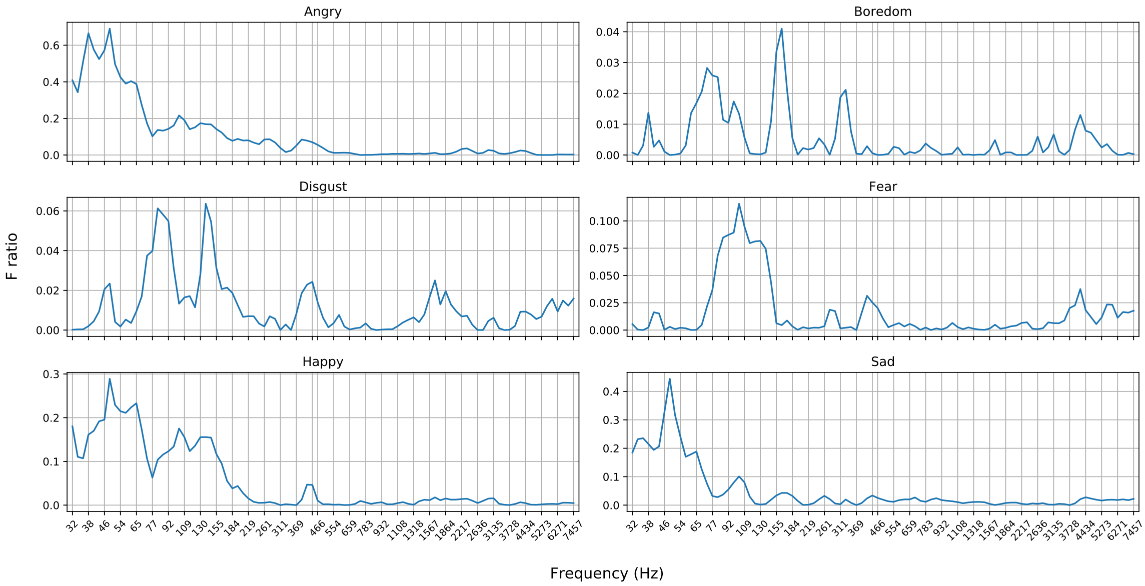


Fig. 2: CQT F-ratio for different frequency bins over EmoDB database.

$$a_k(n) \; = \; \frac{1}{N_k} w\left(\frac{n}{N_k}\right) exp\left[-i2\pi n\frac{f_k}{f_s}\right] \tag{2}$$

where, f_k is the center frequency of a_k, f_s is the sampling frequency, and $w(n)$ is the window function. In this work, *Hann* window is used for CQT calculation. The non-linear placement of f_k in CQT is given by the relation $f_k \; = \; f_{min}2^{\frac{k-1}{B}}$ with B being the number of frequency bins used per octave of frequency and f_{min} is the frequency of the lowest bin. This bin placement is inspired by the equal-temperament scale used in music analysis [44]. The constraints of equal-temperament scale and constant-Q factor of filters result in window length varying over frequency index k, given by,

$$N_k \; = \; \frac{f_s}{f_k}Q. \tag{3}$$

Equation 1 describes the CQT computation as convolution of the *atom* with every input signal sample. However, to reduce redundancy in feature and computational complexity, [55] proposed a CQT implementation which uses hop-length as a parameter to shift the constant-Q atom (or constant-Q basis) by a specific number of samples and hence provides time-frames in the time-frequency representation. Further improvements in CQT computation mentioned in [55] are,

- Computation of frequency-domain inner product (correlation) between the signal frame and time-frequency atom (spectral kernel) instead of time-domain computation.

- Use of only non-zero values of sparse spectral kernels for computation.

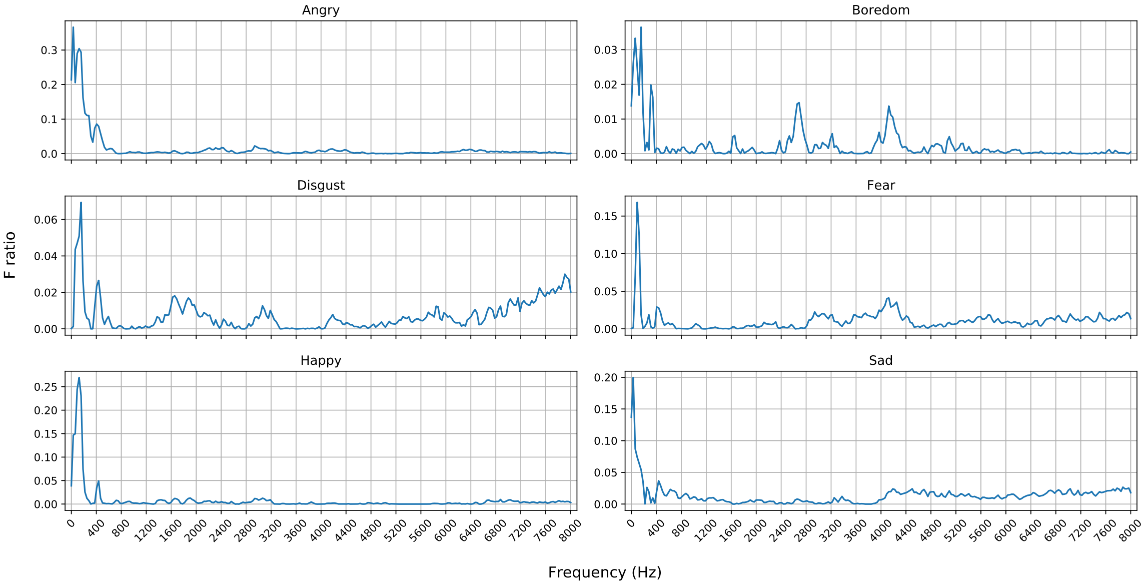


Fig. 3: Spectrogram F-ratio for different frequency bins over EmoDB database.

- Octave-wise transformation, starting from the highest octave followed by down-sampling and low-pass filtering of the signal to obtain lower octaves.

These computation steps differentiate CQT from CWT and provide a non-linear time-frequency representation with reduced redundancy and computational complexity. In our experiments, we used the CQT implementation provided in the *LibROSA*[2] toolbox which uses the above mentioned computational improvements.

When discrete cosine transform (DCT) is applied on CQT values after uniform resampling, the obtained coefficients are called constant-Q cepstral coefficients (CQCC) [44]. However, when DCT is applied without resampling of CQT coefficients, we obtain constant-Q coefficients (CQC) [56]. To get an estimate of class-separability of time-frequency representations, we performed the F-ratio analysis (as given in [57]) on frequency bins obtained from CQT and MFSC. Figure 2 and 5 show the F-ratio statistic of CQT and MFSC obtained at different frequency bins for different emotions with respect to *Neutral* emotion on EmoDB database. The higher F-ratio values at low-frequency bins of CQT and MFSC show the presence of more emotion discriminative information at these bins. The figures also indicate that CQT-spectrogram has higher percentage of discriminative coefficients on an average due to higher resolution in low-frequency regions. Similarly, Fig. 3 shows the F-ratio plot for STFT (short-time Fourier transform) based spectrogram. Although some emotion classes in spectrogram has higher F-ratio values, the percentage of such discriminative coefficients is low in total. However, most of the higher F-ratio coefficients are again observed at low frequencies.

[2] `https://librosa.github.io/`

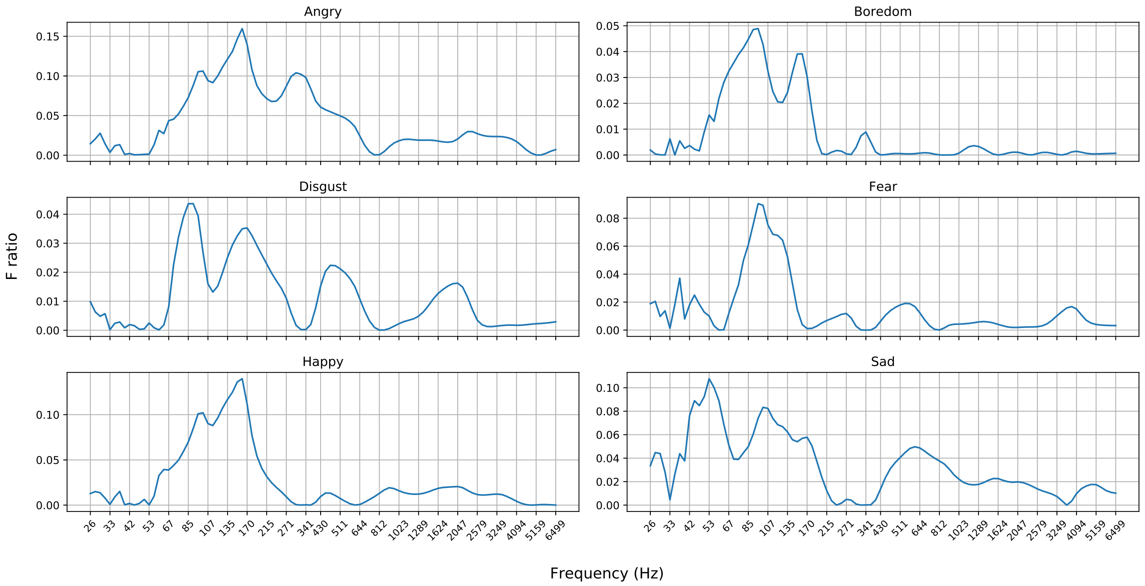


Fig. 4: CWT F-ratio for different frequency bins over EmoDB database.

CQT was first introduced for music analysis owing to its ability to model an equal-tempered frequency scale followed in Western music [44]. Studies performed in psychology also explain some relationship between music training and emotion perception. Individuals with music expertise can better perceive emotions from speech [58, 59]. As CQT is a well-matched representation for music, this could also explain its suitability for SER. Although the CQT configuration found best for SER in our previous contribution ([54]) differs slightly from that used in music analysis, the division of complete frequency range in octaves with equal number of bins supports the similarity between music perception and SER.

3.2. Continuous wavelet transform

CWT is another transform that provides time-frequency representation of the input signal with varying frequency resolution [60]. CWT of a signal $x(t)$ is given as,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\psi\left(\frac{t-b}{a}\right) dt \tag{4}$$

where, $\psi(t)$ is the wavelet basis function while a and b refer to the scale and translation values of the basis, respectively. The large scale values of basis are used to extract slowly varying attributes, whereas smaller scale values can capture minute details of the signal $x(t)$. The collection of wavelet basis with selected scale values also provides a constant-Q factor filterbank representation [60]. Moreover, the scale values in CWT can also be selected to obtain an equal-temperament representation similar to CQT. Thus, CWT also provides high frequency resolution at low frequencies and high time resolution at

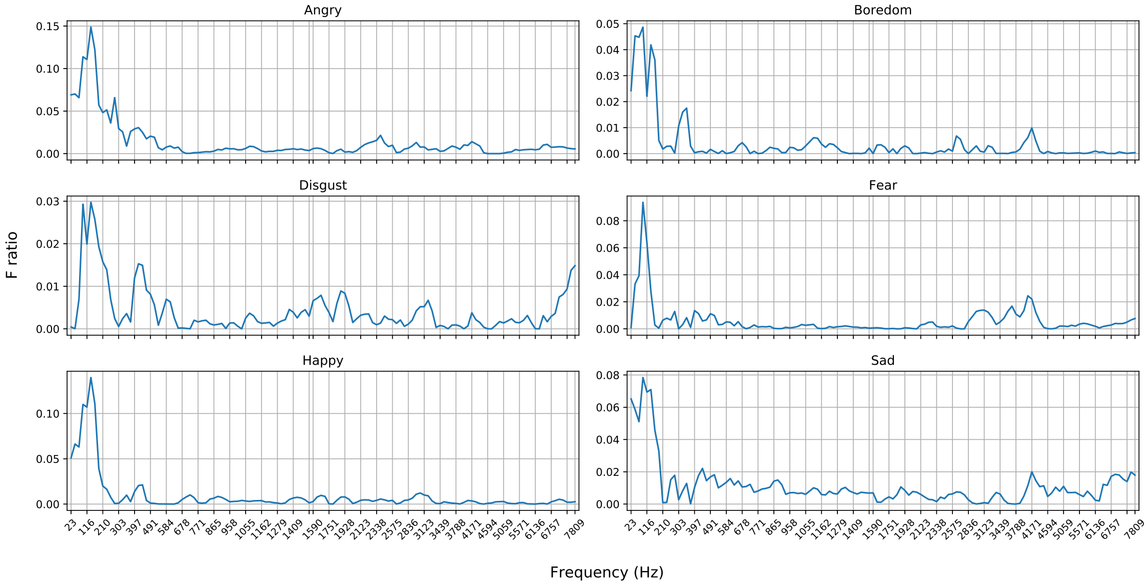


Fig. 5: MFSC F-ratio for different frequency bins over EmoDB database.

high frequencies. As described by Eq. 4, CWT coefficients are computed for every integer value of dilation u leading to a highly redundant and computationally extensive coefficient representation for different scale values. This makes the use of CWT as time-frequency representation less efficient. As the CNN requires the input in image (2-dimension) form, we performed sliding window fashioned summation of CWT coefficients to reduce the size of image (time-frequency representation) in time axis, making it more suitable to be used with CNN. However, the issue of redundancy and computation time remains unaffected. Our coefficient summation approach is also based on windowing performed in MFSC to obtain frames in the corresponding time-frequency representation.

In our CWT implementation, we used *complex Morlet* wavelet pertaining to its relatedness with human perception of vision and sound [53]. We used dyadic scale values (2^a, where $a \in \mathbb{R}$) in CWT for fair comparison with CQT time-frequency representation. The discrete wavelet transform (DWT) provides orthogonal wavelet bases which provide uncorrelated time-frequency features at their disposal. Also, since CNNs are efficient in extracting correlation across time and frequency, redundancy due to correlation among features available in CWT should assist in better capturing of correlation across CWT coefficients for SER. Hence, we exclude DWT from our experiments. Figure 4 shows the F-ratio plot of frequency bins of CWT for different emotion classes in EmoDB database. The plot shows that CWT also has frequency bins which are discriminative (have greater F-ratio values). However, the magnitude of F-ratio values are less for every emotion indicating lesser discriminative characteristics of CWT coefficients.

3.3. Mel-frequency spectral coefficients

Mel-frequency based features, specifically MFSC and MFCC, are widely used in speech signal processing. These features are computed by framing and windowing the speech into short segments followed by computation of power spectrum of every frame [61]. This provides a time-frequency representation with number of time frames on one dimension and number of frequency bins on the other. A mel-frequency based filterbank is then applied on the frequency response of every frame to obtain an output coefficient from each filter and hence introduce perception based non-linearity on the time-frequency representation.

Although the non-linearity in mel-scale is based on human sound perception, it was mainly designed to recognize phonemes in speech. Also, mel filterbank provides poor low-frequency resolution as compared to constant-Q filterbank. As a result, the time-frequency representation generated using mel-scale provides less emphasis on the emotion salient frequency regions of speech signal. The lower F-ratio values of MFSC in Fig. 5 throughout the complete frequency range further indicates the inferiority of MFSC to discriminate among different emotions classes.

We also performed the F-ratio analysis of CQT, CWT, MFSC, and spectrogram feature on IEMOCAP and eNTERFACE. We found higher emotion discriminability of low-frequencies with these databases as well. However, here we report F-ratio analysis only over EmoDB to prevent text redundancy and to have F-ratio plots made with audio files containing the same contextual information (i.e., transcript).

4. Comparison between time-frequency representations

To understand the effect of filters applied on speech with non-linearly spaced center frequencies from SER perspective, we performed both time and frequency domain comparison between mel-scale and constant-Q based time-frequency representations. In time-domain analysis, we analyzed the time-warp stability and time-shift invariance of MFSC by reproducing the analysis performed in [62] and then extended it for analysis on constant-Q filterbank based features.

4.1. Time-domain representation of MFSC

Consider a signal $x(t)$, its time-shifted version is given by $x(t-c)$. The relation between Fourier transform of the original and the shifted version of the signal is,

$$X_c(\omega) \;=\; e^{-j\omega c} X(\omega). \tag{5}$$

Applying modulus on both sides, we obtain,

$$|X_c(\omega)| \;=\; |X(\omega)|. \tag{6}$$

Equation 6 shows that the modulus of the Fourier transform remains stable to time shift c under the condition that $c \;<<\; T$, where T is the duration of the window over which the Fourier transform is computed. Hence, Fourier transform is inherently stable but only for duration T.

Consider a time-warped (deformed) version of the signal given by $x_\tau(t) \;=\; x(t-\epsilon t)$, with $0 < \epsilon \ll 1$. The modulus of Fourier transform of the time warped signal is given as,

$$|X_\tau(\omega)| = \left| \frac{1}{1-\epsilon} X\left(\frac{\omega}{1-\epsilon}\right) \right|. \tag{7}$$

The factor $\frac{1}{1-\epsilon}$ in the frequency term leads to a shift of the frequency components at ω' by a factor $\epsilon|\omega'|$. This effect gets more prominent at higher values of ω. Hence, the Fourier transform is stable to time-shifts but not to time-warping. Such effect would cause the same emotion utterance of different speakers (e.g., same utterance of a male and a female speaker) to have different frequency attributes making the classification process more difficult. Also, in STFT, the utterance is divided into frames of smaller duration (for example, 20 ms) and then the modulus of Fourier transform is computed for every frame. The STFT, therefore, remains invariant to time-shifts existing only under 20 ms duration. However, since speech segments with length greater than 250 ms contain the information required for emotion prediction [17], the STFT fails to capture it efficiently. Therefore, the STFT (spectrogram) time-frequency representation is not an effectual representation from SER perspective.

In mel-scale based representations, a mel-filterbank is applied on the computed STFT to obtain MFSC/MFCC. The mel-filterbank contains filters with logarithmically spaced center frequencies. Let the signal frame centered at time t be given as $x_t(u) = x(u)\phi(u-t)$, where u is the time index and ϕ is the framing window. In MFSC, the Fourier transform of the signal frame ($X_t(\omega)$) is multiplied with Fourier transform of every mel-filter ($\psi_\lambda(\omega)$) and then summed to obtain a single coefficient at the output of every filter. Mathematically, this is given as,

$$Mx(t,\lambda) = \frac{1}{2\pi} \int |X_t(\omega)|^2 |\psi_\lambda(\omega)|^2 d\omega \tag{8}$$

where, λ is the support or center frequency of the mel-filter, t is the center of the time frame and $Mx(t,\lambda)$ is the corresponding MFSC . Equation 8 can also be considered as averaging in frequency domain. Converting the multiplication in frequency-domain in Eq. 8 into convolution in time-domain, we get,

$$\begin{aligned} Mx(t,\lambda) &= \int |x_t * \psi_\lambda(v)|^2 dv \\ &= \int \left| \int x(u)\phi(u-t)\psi_\lambda(v-u)du \right|^2 dv. \end{aligned}$$

Since the filter support λ of mel-filters in time is generally smaller than the support of ϕ, we have,

$$\begin{aligned} Mx(t,\lambda) &\approx \int \left| \int x(u)\psi_\lambda(v-u)du \right|^2 |\phi(v-t)|^2 dv \\ &= |x * \psi_\lambda|^2 * |\phi|^2(t). \end{aligned} \tag{9}$$

Equation 9 shows that averaging performed in frequency-domain (Eq. 8) is equivalent to time-domain averaging by filter ϕ. Also, the averaging in frequency domain in Eq. 8 provides one single coefficient representation of a band of frequencies (defined by bandwidth of mel-filters). This averaging makes the MFSC representation less susceptible to frequency shifts and hence stable to time-warps [62]. Thus the MFSC representation is

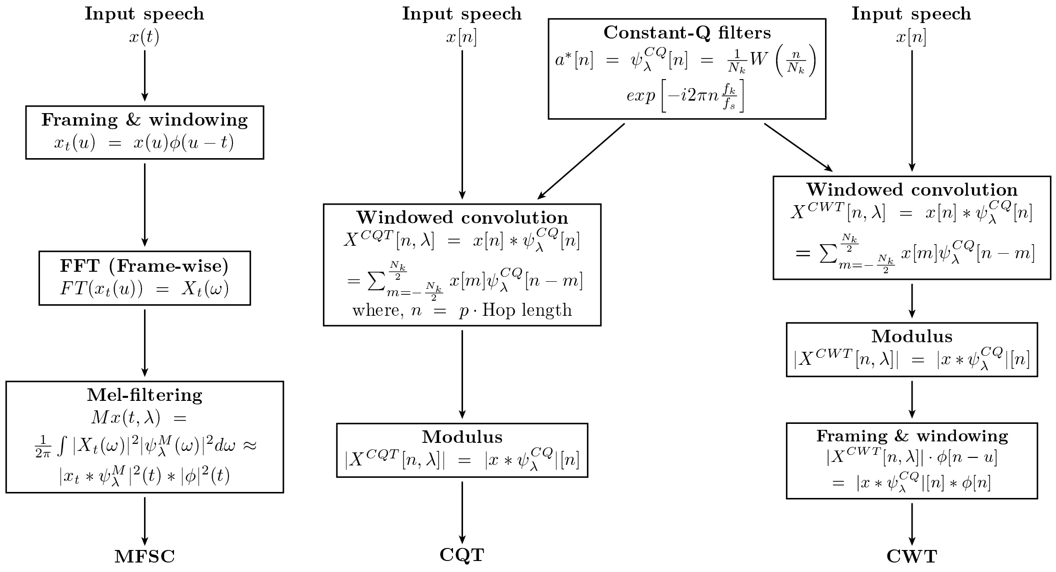

Input speech
$x(t)$
Framing & windowing
$x_t(u) = x(u)\phi(u-t)$
FFT (Frame-wise)
$FT(x_t(u)) = X_t(\omega)$
Mel-filtering
$Mx(t,\lambda) = \frac{1}{2\pi}\int |X_t(\omega)|^2|\psi_\lambda^M(\omega)|^2 d\omega \approx |x_t * \psi_\lambda^M|^2(t) * |\phi|^2(t)$
MFSC
Input speech
$x[n]$
Constant-Q filters
$a^*[n] = \psi_\lambda^{CQ}[n] = \frac{1}{N_k} W\left(\frac{n}{N_k}\right) exp\left[-i2\pi n \frac{f_k}{f_s}\right]$
Windowed convolution
$X^{CQT}[n,\lambda] = x[n] * \psi_\lambda^{CQ}[n] = \sum_{m=-\frac{N_k}{2}}^{\frac{N_k}{2}} x[m]\psi_\lambda^{CQ}[n-m]$
where, $n = p \cdot$ Hop length
Modulus
$|X^{CQT}[n,\lambda]| = |x * \psi_\lambda^{CQ}|[n]$
CQT
Input speech
$x[n]$
Windowed convolution
$X^{CWT}[n,\lambda] = x[n] * \psi_\lambda^{CQ}[n] = \sum_{m=-\frac{N_k}{2}}^{\frac{N_k}{2}} x[m]\psi_\lambda^{CQ}[n-m]$
Modulus
$|X^{CWT}[n,\lambda]| = |x * \psi_\lambda^{CQ}|[n]$
Framing & windowing
$|X^{CWT}[n,\lambda]| \cdot \phi[n-u] = |x * \psi_\lambda^{CQ}|[n] * \phi[n]$
CWT

Fig. 6: Block diagrams showing time-domain formulation of different features used in this work. Here, ψ_λ^M and ψ_λ^{CQ} represent mel and constant-Q filters with frequency support λ, $\phi(t)$ is the averaging window, $p \in \mathbb{Z}^+$, N_k and f_k are the time-span and frequency of the k^{th} constant-Q basis, and $*$ denotes convolution operation. We show the MFSC feature extraction process in continuous time-domain representation (t) for similarity with analysis performed in [62].

invariant to both deformation and time-shift but only for a specific frame duration (here, 20 ms). Therefore, although stable, MFSC is still incapable of capturing long time scale information, which is essential for SER [17]. The left most column of Fig. 6 shows a pictorial representation of different steps involved in MFSC computation.

4.2. Time domain representation of CQT

We now show similar analysis on CQT to investigate the time and deformation invariance of CQT. From Eq. 1, the time-shifted version of CQT is given as,

$$X^{\text{CQT}}[k, n-c] \;=\; \sum_{p=(n-c)-\lfloor N_k/2 \rfloor}^{(n-c)+\lfloor N_k/2 \rfloor} x(p) a_k^*(p-(n-c)+N_k/2) \tag{10}$$

where, k and n are the frequency and time bins respectively, N_k is the length of k^{th} basis function, c is the time shift, X^{CQT} is the corresponding CQT coefficient, and,

$$a_k^*(n-c) \;=\; \frac{1}{N_k} W\left(\frac{n-c}{N_k}\right) exp\left[-i2\pi(n-c)\frac{f_k}{f_s}\right]. \tag{11}$$

Therefore, the relation between CQT of time shifted and original version of signal x is given as,

$$X^{\text{CQT}}[k, n-c] \;=\; exp\left[-i2\pi c\frac{f_k}{f_s}\right] X^{\text{CQT}}[k, n]. \tag{12}$$

Again applying modulus on both sides,

$$\left|X^{\mathrm{CQT}}[k, n-c]\right| = \left|X^{\mathrm{CQT}}[k, n]\right|. \tag{13}$$

We observe that the modulus of CQT exhibits the same time-shift invariance as observed in STFT. However, in CQT, the time window over which the transform is computed is *not fixed.* The time window duration varies with the relation $N_k = Q\frac{f_s}{f_k}$. Therefore, the time-invariance of CQT also varies with the frequency bin k. Since the value of N_k is higher for small values of k and vice-versa, the time-invariance varies from high to low towards high frequencies in the CQT representation. The same is also evident from the smearing of frequency components observed in Fig. 7. This is in contrast with the fixed time-invariance in MFSC, defined by the duration of window function $\phi(t)$. This property of CQT should help provide more time-invariance to the emotion relevant characteristic of speech, i.e., the pitch frequency [34, 35]. With higher invariance the effect of emotion irrelevant variations in pitch, e.g., variations due to different speaking styles, contexts, etc., can be reduced, hence improving the complete emotion representation. At the same time, the quick variations in time that takes place at high frequency can also be simultaneously captured due to higher time-resolution. This property is more useful for emotions with comparatively more energy at high frequencies, for example, *Angry* and *Happy.* The Eq. 9 equivalent of CQT can be given as,

$$X^{\mathrm{CQT}} = |x * \psi_\lambda|(t). \tag{14}$$

Hence, the time-frequency representation of CQT is the modulus of convolution of signal x with different filters (ψ_λ) in CQT filterbank. The modulus term provides time-invariance to filterbank convolution output. The middle column of Fig. 6 shows different steps in CQT computation. Regarding stability to time-warping deformations, CQT also employs varying bandwidth filters with same effect as averaging across different frequency bands. This introduces similar deformation stability as done by mel-filter bank in MFSC. Figure 8 explains the effects of deformations on STFT and CQT and the corresponding stability in CQT. The stability removes the variations appearing due to different speaking styles, e.g., slower or faster speaking rate of different individuals.

4.3. Comparison between CQT and CWT

CWT also provides a time-frequency representation generated from a constant-Q filterbank [60]. However, the CWT representation is redundant as it does not include downsampling of band-pass filter responses. Also, due to its larger size, it is difficult to consider raw CWT as input to CNN. To alleviate this, our employed implementation includes CWT computation followed by framing and summation of the coefficients. The Eq. 9 equivalent of CWT computation is given as,

$$X = x * \psi_\lambda(t) \tag{15}$$

where, ψ_λ is band-pass wavelet basis with support at λ. The framing and averaging performed on computed CWT can also be given as low-pass filtering of Eq. 15. Hence,

$$X^{\mathrm{CWT}} = |x * \psi_\lambda| * \phi(t), \tag{16}$$

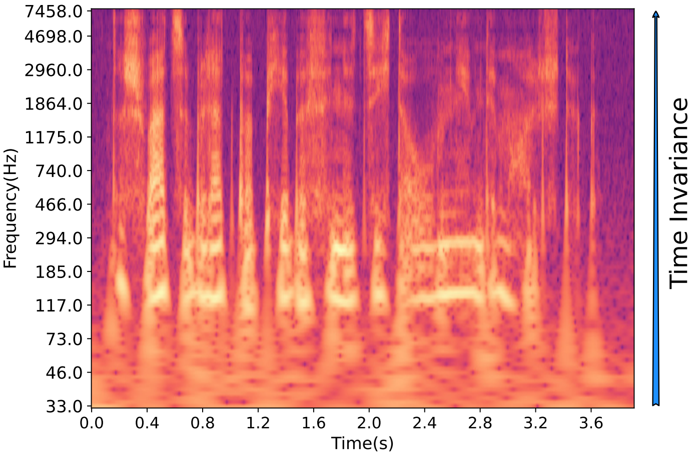
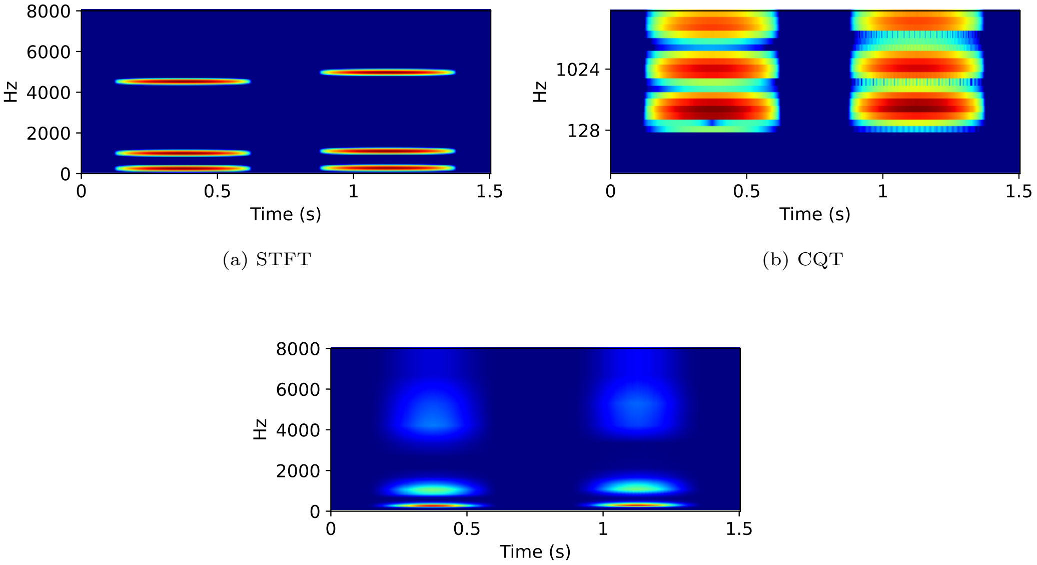


Fig. 7: The frequency varying time invariance in CQT. Due to higher value of N_k, smearing in time domain is visible at low frequencies.

where, ϕ is the framing window and X^{CWT} is the corresponding CWT coefficient. The modulus is applied to again remove the phase component and obtain time-invariant response. The right most column of Fig. 6 describes the computation of CWT. As the filters used in CWT closely follow the structure of CQT filters, Eq. 15 also provides varying time-invariance and higher low-frequency resolution. However, the averaging in Eq. 16 fixes the extent of time-invariance, similar to MFSC. In our experiments, CWT's frame duration is also kept equal to 20 ms following MFSC. Therefore, the time invariance in X^{CWT} varies for low-frequency bins, i.e., for filters with time span greater than 20 ms, and remain fixed to 20 ms for filters with span less than 20 ms. Eq. 16 closely follows the time-domain MFSC (Eq. 9) and the first layer scattering domain coefficients explained in [62] for 1D signals. The comparison of CQT with the CWT time-frequency representation used in this work would help to analyze the importance of varying time-invariance and improved low-frequency resolution, as compared to standard MFSC. This is further consolidated with the experiments that follow in later sections.

4.4. Frequency domain comparison of mel and constant-Q filterbanks

The major difference between the MFSC and CQT time-frequency representation in time domain appears in the form of fixed and varying time invariance across different frequency bins. In frequency domain, the major difference exists in the center frequencies of filters in the filterbank. Mel-scale follows a decadic logarithm scale, whereas CQT follows a binary logarithm scale. This leads to higher low-frequency resolution in CQT as compared to MFSC. Another property of the log-frequency is that the distance between pitch harmonics is invariant to pitch frequency. This is contrast with in linear frequency representation (e.g., spectrogram) where the harmonic distance reduces with reduced pitch frequency. Figure 9 describes this phenomenon. This also helps CQT (or constant-Q response, in general) in better resolution of pitch and its harmonics as compared to



(a) STFT

(b) CQT

(c) CQT with linear frequency scale

Fig. 8: Comparison between deformation stability of STFT, CQT, and CQT with linear frequency scale. The figures include frequency response of a synthetic signal $x(t)$ consisting of three tones of $0.5s$ duration (on left) and its deformed counter-part $x'(t) = x((1-\epsilon)t)$ (on right). $x(t)$ is made by adding three tones of 250 Hz, 1000 Hz, and 4500 Hz frequency. In a), harmonics with high frequency get more severely affected by the distortion. In b), due to filters with high bandwidth, the harmonics in the original ($x(t)$) and deformed signals ($x'(t)$) overlap significantly, hence reducing the effect of distortion at high frequencies. In c), we show the linear frequency scale response of $x(t)$ and $x'(t)$ with CQT. The varying bandwidth of filters leads the CQT representation to inherent stability against time-warp deformation.

STFT or MFSC. For further analysis, we compute the mean and standard deviation of pitch and first pitch harmonic, averaged across every utterance for different emotions of male and female speakers in EmoDB database. We then generate tones corresponding to average pitch and its first harmonic frequency and compute the CQT and MFSC representation of the tones. Figure 10 shows the obtained representations. We observe that CQT better resolves the pitch and its first harmonic as compared to MFSC for both male and female speakers. MFSC shows overlapping between pitch and its harmonic for emotions with low mean pitch frequency (*Boredom*, *Neutral* and *Sadness*). However, CQT clearly differentiates pitch and its harmonics for every emotion in EmoDB. Also, for *Disgust*, *Neutral* and *Sad* emotions, the separation in CQT representation of females is higher than that in males. This is because of the naturally higher pitch of female speakers as compared to that of the males. The error bars show the standard deviation of pitch for various emotions in EmoDB. Same value of standard deviation show different dynamic range over frequency bins in CQT and MFSC. This emphasizes the difference in low-frequency resolution in MFSC and CQT and also the superior pitch resolution in CQT.

The studies in [63] and [64] suggest that a filterbank structure with high resolution (dense filters) on mid and high frequency regions are beneficial for speaker recognition.

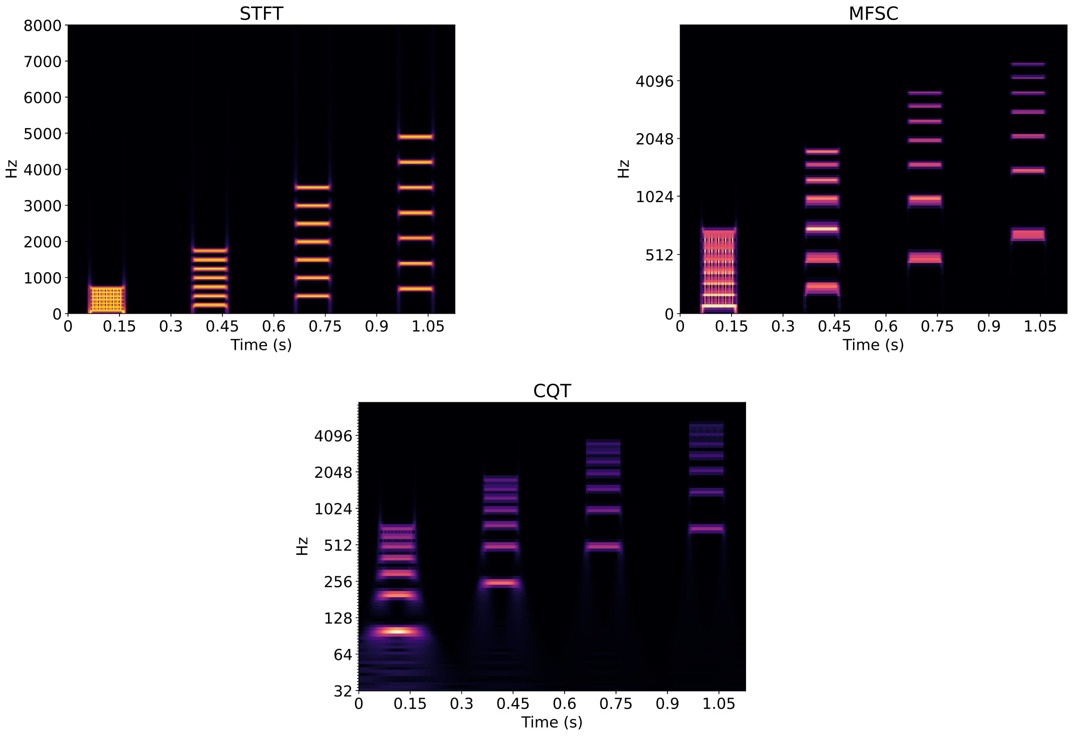


Fig. 9: Time-frequency representation of signal consisting of 100 Hz, 250 Hz, 500 Hz, and 700 Hz frequency tones with corresponding first six harmonics. The complete signal is made by concatenating the sub-sequences (tones with corresponding harmonics) with each other. In STFT, the distance between harmonics increases with increase in fundamental frequency (pitch). For log-frequency based scale (mel-scale and constant-Q scale), the distance between harmonics should remain invariant to pitch frequency. In mel-scale, due to the decadic logarithm scale, the distance varies for low pitch values. For CQT, distance remains unchanged for every pitch value.

In contrast to this, the dense filter arrangement at low-frequency in CQT and CWT should provide speaker complimentary information. Therefore, constant-Q representation should also remain invariant to speaker information compared to MFSC (because of more low-frequency resolution) facilitating SER.

5. Experimental setup

5.1. Neural network architectures

In this subsection, we briefly review the different deep neural network architectures that were employed to evaluate SER performance of features. Our choice of these architectures was inspired by the success of techniques such as 1D and 2D convolutions, LSTM, attention mechanism, squeeze and excitation module, Res2Net module, etc. in different speech processing domains [15, 29, 65–68].

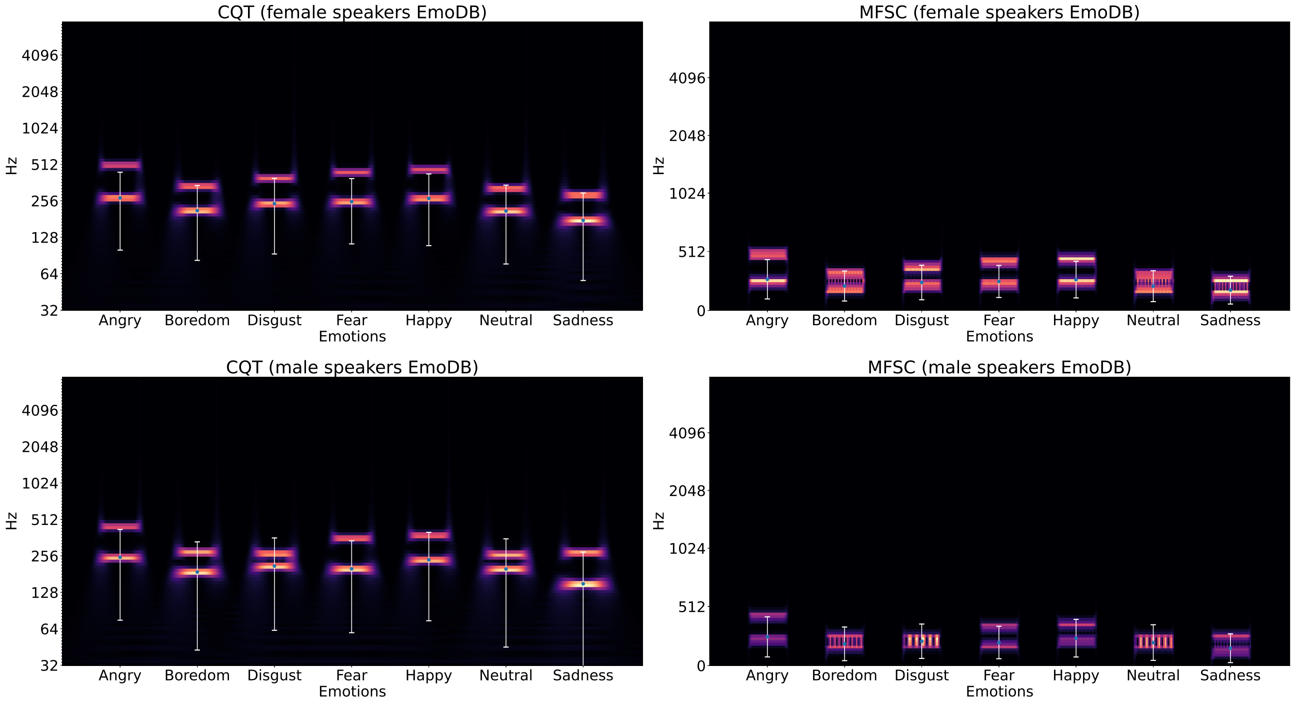


Fig. 10: Pitch and first pitch harmonic of Female and Male speakers for various emotions in EmoDB. In the figure, for every emotion, the lower stripe shows the mean pitch averaged over all utterances in EmoDB. The upper stripe shows the mean first pitch harmonic averaged over EmoDB. The error bars show the standard deviation of pitch values for different emotions. Due to higher low-frequency resolution, CQT can better resolve and differentiate between pitch and its first harmonic across both female and male speakers.

5.1.1. TDNN architecture

As a 1D convolutional network, a TDNN architecture, shown in Table 1, was used for SER with the time-frequency representations as inputs. The TDNN structure was inspired by the *x-vector* systems developed for speaker verification [69]. A TDNN structure processes frame-level information in the convolutional layer by applying different dilation values at different layers to efficiently capture the temporal spread of information. This is followed by a statistics pooling layer which aggregates the processed information in the temporal dimension to generate a segment level representation. The segment-level features are then fed to FC layers and a softmax layer for final classification. The TDNN structure used in this work employed an end-to-end classification system, in contrast with the structure used in [69], which extracts segment-level speaker embeddings from the statistics pooling layers for classification with a different classifier.

5.1.2. Conv2D architecture

Table 1 also shows the architecture of a 2-dimensional CNN or Conv2D network used for SER classification. Conv2D network involves convolution of the input feature matrix with a kernel which allows movement in both time and frequency dimensions of input time-frequency representation. This helps in capturing the feature correlation across both

Table 1: Parameters of TDNN, Conv2D, and Conv2D-LSTM architecture used in this work.

Architecture	Layers	No. of Filters	Height (Frequency)	Length (Time)	Dilation
TDNN	1D Conv	32	5	5	1
	1D Conv	32	3	3	2
	1D Conv	32	3	3	3
	1D Conv	64	1	1	1
	Mean & Standard Deviation Pool	-	-	-	-
	Fully Connected	64	-	-	-
	Softmax	#Classes	-	-	-
Conv2D	2D Conv	32	5	5	1
	MaxPool	-	2	1	1
	2D Conv	32	3	3	1
	2D Conv	32	3	3	1
	2D Conv	64	1	1	1
	Mean & Standard Deviation Pool (over time)	-	-	-	-
	Flatten	-	-	-	-
	Fully Connected	64	-	-	-
	Softmax	#Classes	-	-	-
Conv2D-LSTM	2D Conv	64	3	5	1
	MaxPool	-	2	2	1
	2D Conv	64	3	5	1
	MaxPool	-	2	2	1
	2D Conv	64	3	5	1
	MaxPool	-	2	2	1
	2D Conv	64	3	5	1
	MaxPool	-	2	2	1
	2D Conv	64	3	5	1
	2D Conv	64	3	5	1
	Mean Pool (over frequency)	-	-	-	-
	BLSTM	64	-	-	-
	Fully Connected	64	-	-	-
	Softmax	#Classes	-	-	

time and frequency dimensions. As the emotion information is known to remain temporally spread throughout the utterance, 2D convolution windows with different kernel sizes helps in information extraction across different time scales. Similar to the TDNN architecture, our Conv2D architecture also applied temporal mean and standard deviation pooling on the output of the final convolution layer. This operation again aggregates the features extracted by convolutional layers across time dimension to obtain a vector representation of the input segment. The temporal pooling is followed by FC and Softmax layers for classification. We used the Conv2D network for end-to-end classification, with speech features at the input and probable emotion class at the output. We kept the parameters, i.e., no. of filters, kernel sizes, and number of layers in Conv2D same as used in TDNN architecture for comparison. We also placed a max-pooling layer after the first convolutional layer to reduce parameters in Conv2D architecture and prevent overfitting.

5.1.3. MERC2020 baseline architecture

To compare SER performance across different architectures, we used the speech emotion recognition model proposed in multimodal emotion recognition competition 2020 (MERC 2020). The model contains three Conv2D layers, one LSTM layer followed by attention pooling and a dense layer. We used the implementation provided by the organizers of the MERC2020[3].

5.1.4. Attention-based LSTM

We also evaluated the performance of different features with attention-based LSTM model proposed in [65]. The model includes a modified LSTM, in which the forget gate is replaced with an attention mechanism called attention gate. The attention gate provides increased focus on the most emotion-relevant parts of the input time-frequency representation. Moreover, this modification decreases the number of trainable parameters in every attention-based LSTM block, causing a reduction in the model training time [65]. The output of attention-based LSTM is fed to two separate attention layers, one focusing on time and the other on the frequency dimension. Hence, the complete architecture consists of two layers of modified LSTM units (attention-based LSTM blocks) followed by parallelly placed time and frequency attention layers. The activations from attention layers are concatenated and fed to two FC layers and softmax for final classification. We chose the same parameter values for different layers as used in the original paper.

5.1.5. Transformer encoder model

We employed a transformer encoder model proposed in [66] to compare with the state-of-the-art deep learning architectures. In [66], information from various modalities (speech, speaker, and text) was extracted using encoder blocks, and the output was concatenated for SER. Since our approach includes only speech modality, we chose the speech signal's transformer encoder block from [66] in our experiments. The model consists of a Conv1D layer followed by a combination of multi-head attention and linear layers, constituting the encoder block. The output of the encoder block is then fed to a frame-level (or time-level) attention pooling layer followed by an FC and softmax. The parameter values of the neural network layers were the same as used in the original paper.

5.1.6. ECAPA-TDNN

The emphasized channel attention, propagation, and aggregation in time-delay neural network (ECAPA-TDNN) proposed in [68] introduces multiple enhancements over the standard x-vector TDNN architecture [69]. These include using multiple 1D Res2Net modules, squeeze and excitation blocks, and channel-dependent time-frame attention. The ECAPA-TDNN architecture is found useful for speaker recognition and speaker diarization tasks [67]. In this work, we used the implementation of ECAPA-TDNN provided in SpeechBrain[4] Python toolkit without any change in parameter configuration.

[3] https://github.com/ki4ai-skc/merc2020

[4] https://speechbrain.github.io/

5.1.7. Conv2D-LSTM

In [70], an attention-based convolutional recurrent neural network (ACRNN) was proposed for SER. Inspired by this, we designed an architecture consisting of only CNN and LSTM layers. We removed the attention layer from this architecture to analyze the effect of the temporal memory-based recurrent layer on CNN learned features and compare it with plain Conv2D architecture. Also, the MERC2020 model already included a combination of LSTM and attention with convolution layers. Table 1 describes our employed Conv2D-LSTM architecture. An LSTM contains a memory element that can accumulate the required information across several time frames. Using LSTM on CNN activations helps extract emotion-relevant temporal information from translation invariant and downsampled feature representations provided by CNN. Since the last time-frame output of the LSTM contains the most abundant emotion information [65], we fed only this output to the final FC and softmax layers and discarded the remaining time frames. This architecture also helped to compare the temporal aggregation of emotion information of LSTM with that of attention layers in Attention-based LSTM, Transformer encoder, and ECAPA-TDNN architectures.

For the above mentioned deep networks, we used *Keras*[5] deep learning library for Conv2D, TDNN, and Conv2D-LSTM architecture and *PyTorch*[6] deep learning library for the remaining selected state-of-the-art architectures.

5.2. Databases

We used four different databases for analysis and evaluation of features. They are freely available and widely used in SER. Performance of selected features on these corpora also facilitates comparison with similar SER methods. Table 2 summarizes the different SER databases used in our experiments.

5.2.1. Berlin emotion database (EmoDB)

Berlin emotion database (EmoDB) [71] is one of the most widely used database in SER. It includes acted spoken utterances of 10 professional artists (5 female and 5 male). Ten sentences, emotionally neutral and phonetically rich, are spoken by the actors in German language. The database contains speech recordings of seven different emotions: *Angry, Happy, Fear, Sad, Boredom, Disgust* and *Neutral*. The authenticity of the recorded emotions was evaluated by listening test performed on 20 subjects. In total, the actors recorded 800 utterances but only 535, having more than 80% recognition rate and 60% naturalness, were finally selected. The mean recognition accuracy of emotions in the listening test was 84.3% on the selected 535 recordings. The diligent recording setup and free availability has led this database to be used in various important works [10, 16, 17, 25–27, 37, 51] and hence is used here as well.

5.2.2. Ryerson audio-visual database of emotional speech and song (RAVDESS)

The RAVDESS database [72] contains emotion speech and song samples recorded from 12 male and 12 female artists speaking English language. The database contains

[5] `https://keras.io/`

[6] `https://pytorch.org/`

Table 2: Summary of SER Databases.

Databases	Type	Speakers	Emotions	Sampling Rate	Total Utterances	Language
Berlin Emotion Database (EmoDB) [71]	Acted	10 (5 Female) (5 Male)	7 (Anger, Sad, Boredom, Anxiety/Fear, Happy, Disgust and Neutral)	16 kHz	535	German
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [72]	Induced	24 (12 Female) (12 Male)	8 (Calm, Happy, Sad, Anger, Neutral, Fearful, Surprise, and Disgust)	48 kHz	1440	English
Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [32]	Induced	10 (5 Female) (5 Male)	4 (Happy, Sad, Anger and Neutral)	16 kHz	4936	English
eNTERFACE '05 [73]	Induced	44 (8 Female) (36 Male)	6 (Happy, Sad, Anger, Fear, Surprise and Disgust)	48 kHz	1293	English (Different nationality)

a total of 7536 clips with data recorded in three modalities: audio-only, video-only, and audio-video. The audio-only modality contains 1440 speech utterances from all speakers spoken with eight different emotions (*Happy, Angry, Sad, Neutral, Disgust, Calm, Surprised* and *Fear*) and two intensity levels, strong and normal. Evaluation of recorded clips were performed by 319 subjects, out of which 247 evaluated the validity and 72 provided test-retest reliability of recorded emotions. An average of 60% accuracy was obtained in validity test on recordings of all emotions. Recent design and inclusion of an extensive emotion set with varying intensities make this an important database for SER.

5.2.3. Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)

IEMOCAP is an audio-visual database recorded on 10 professionally trained English speakers (5 male and 5 female) with two recording methods, scripted and improvised [32]. This makes the IEMOCAP recordings more natural compared to the two databases mentioned previously. Eight emotions (*Happy, Angry, Sad, Neutral, Fear, Disgust, Excitement* and *Surprise*) were captured over a total of 10039 recorded samples (5255 scripted and 4784 spontaneous) with an average utterance length of 4 seconds. Samples were annotated into both discrete and continuous emotion labels by six evaluators. In our work, we used discrete emotion labels from only four classes (*Happy, Angry, Sad* and *Neutral*) of both scripted and improvised recordings because of comparatively sparse speech samples for the remaining emotions and also for better comparison with existing SER literature [20, 30, 31, 74, 75].

5.2.4. eNTERFACE '05

eNTERFACE is also an audio-visual database containing recording of six different emotions: *Happy, Sad, Surprise, Anger* and *Fear*, recorded from 44 different subjects [73]. Although the subjects were from different nationalities, English was the common language for recording the data. This introduced accent variability on recorded samples, representing the real-world scenario in a better way. To induce emotions, subjects were made to read a short story before recording their reactions to the story on a fixed set

Table 3: Optimized parameter settings for different features (Q = Q-factor of the filter, f_s = sampling frequency, $f_k = k^{\text{th}}$ frequency bin). The parameters values are taken from experiments performed in [54].

<table>
<tr><th colspan="5"><b>Mel-frequency features
(MFSC)</b></th></tr>
<tr><td>Library used</td><td>Window length
(samples)</td><td>Hop length
(samples)</td><td>No. of
FFT points</td><td>No. of
filters</td></tr>
<tr><td><i>LibROSA</i>(v0.7.1)</td><td>320</td><td>64</td><td>512</td><td>24</td></tr>
<tr><td colspan="5"><b>Constant-Q transform
(CQT)</b></td></tr>
<tr><td>Library used</td><td>Window length
(samples)</td><td>Hop length
(samples)</td><td>Bins per
octave</td><td>No. of
bins</td></tr>
<tr><td><i>LibROSA</i>(v0.7.1)</td><td>Variable $(Q\frac{fs}{fk})$</td><td>64</td><td>3</td><td>24</td></tr>
<tr><td colspan="5"><b>Wavelet transform
(CWT)</b></td></tr>
<tr><td>Library used</td><td>Window length
(samples)</td><td>Hop length
(samples)</td><td>CWT scale values
and wavelet type</td><td>No. of
bins</td></tr>
<tr><td><i>PyWavelets</i>(v1.1.1)</td><td>320</td><td>64</td><td>$2^{\frac{k}{3}} : k \in [3:26]$
Complex Morlet</td><td>24</td></tr>
</table>

of answers. This introduced genuineness into the subject's reactions. We used only the audio modality having a total of 1293 utterances across all the subjects.

5.3. Training/Testing evaluation

Leave-one-speaker-out (LOSO) cross-validation strategy was employed for evaluation of features with DNN classifiers. The databases were divided into training, validation, and test partitions such that the training and validation groups contained sets of disjoint speakers with one left-out speaker kept for testing. The validation set contained speech utterances of only two speakers. Hence, the number of training-validation-testing sets were same as the total number of speakers for every database. According to the literature, speaker-dependent SER generally fairs better than speaker-independent SER [76]. However, using speaker-independent sets from the database eliminates the chances of the trained classifier being biased towards a set of speakers, and also simulates the practical/real-world scenario. We used LOSO even for RAVDESS and eNTERFACE databases which have higher number of speakers (> 10), unlike the leave-one-speaker-group-out (LOSGO) method used in other works, e.g., [17, 77].

Although LOSO cross-validation is computationally extensive, we can safely ignore the increase in complexity due to the small sizes of available SER databases. Also, using only one speaker for testing provides the advantage of more data samples in training and validation, which is essential for small databases.

We used an energy-based speech activity detector to remove silence parts of speech utterances before feature extraction. We also employed cepstral mean variance normalization on features before providing them to the classifiers. To increase the size of available training data, we employed five-fold data augmentation using additive and re-

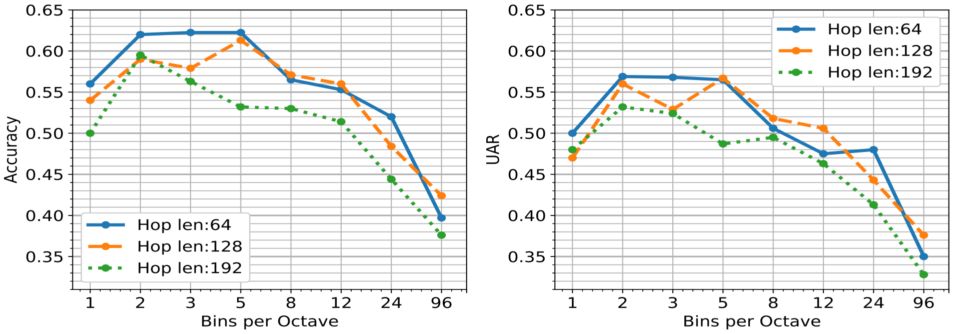


Fig. 11: CQT parameter comparison on EmoDB database. The left subplot shows the change in accuracy with different values of frequency bins per octave. Similarly, the right subplot shows change in UAR with frequency bins per octaves. The total number of octaves are kept fixed at 8. Figure taken from [54].

verberation noises following x-vector extraction recipe [69] used in *Kaldi*[7] toolkit. All available databases were downsampled to 16 kHz before feature extraction.

5.4. Feature evaluation

Table 3 shows the values of the arguments of the built-in feature extraction functions. *LibROSA*[8] *Python library* was used for CQT and MFSC while *PyWavelets*[9] was employed for CWT feature extraction in this work. The parameter values were inspired by our preliminary study, where we performed the optimization of bins per octave and hop length over 8 octaves on EmoDB database [54]. For a fair comparison, similar optimization on MFSC and CWT is employed in this work. We used 24 mel-filter bank with 64 samples hop across different time frames. This corresponds to the optimized CQT parameters values which provided the best results (3 bins per octave with a total 8 octaves, refer Fig 11). In CWT, we used scale values obtained from the expression $2^{\frac{k}{3}}$ where k varies from 3 to 26, to obtain the same number of frequency bins as in CQT. The 3 in $2^{\frac{k}{3}}$ corresponds to the voices per octave with number of octaves again fixed to 8. We chose default values for the remaining input parameters of CQT and MFSC functions in *LibROSA* and CWT in *PyWavelets*.

5.5. Classifier evaluation

For training of DNN architectures, we used non-overlapping segments of input time-frequency representation of length 100 frames. However, during testing, we used the utterances' complete duration. These settings improved the final recognition accuracy compared to the scenario where the test data was truncated. We used a learning rate value of 0.001 with a batch size of 64. The networks were trained for 50 epochs, and

[7] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2
[8] https://librosa.github.io/
[9] https://github.com/PyWavelets/pywt

the model with the highest validation unweighted average recall (UAR) was used during inference. In TDNN, Conv2D, and Conv2D-LSTM architectures, a drop-out value of 0.3 was used in fully connected layers. ReLU activation was used in every layer except for the final softmax layer.

For SER evaluation, we employed commonly used accuracy and UAR as performance metrics. Accuracy is calculated by finding the ratio between the number of correctly classified utterances to the total number of utterances in test set. The UAR metric is given as [78]:

$$\text{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K} A_{ij}} \tag{17}$$

where, A refers to the contingency matrix, A_{ij} corresponds to the number of samples in class i classified into class j, and K is the total number of classes. As accuracy is considered *unintuitive* for databases with unequal samples across different classes, we optimized the feature extraction parameters based on the UAR metric.

6. Results and discussion

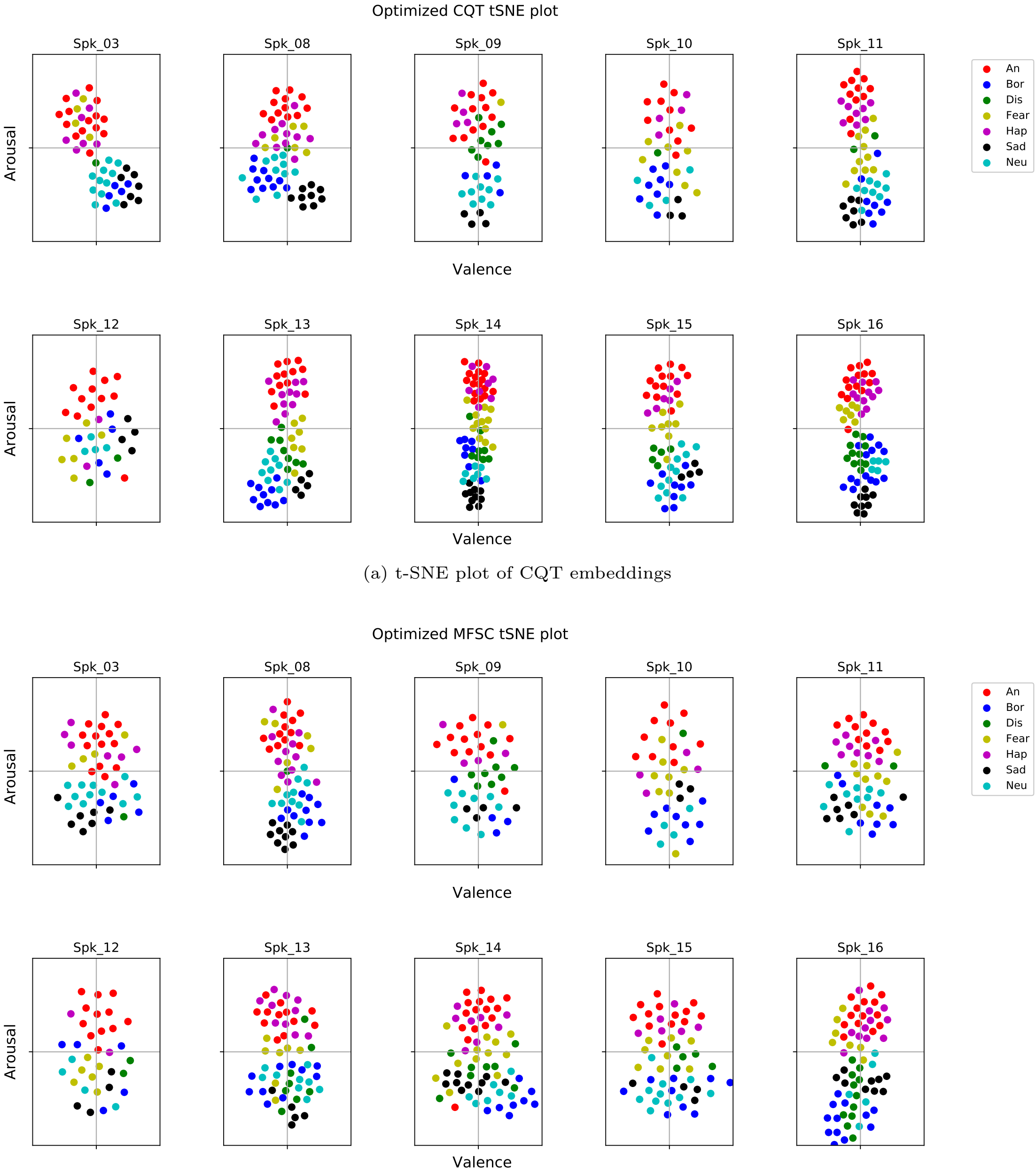
6.1. Embedding visualization

To understand the discriminability among the constant-Q and mel-scaled representations, we present the tSNE plots of the speech embeddings of CQT and MFSC extracted from the statistics pooling layer of the TDNN architecture with EmoDB database in Fig. 12. We kept the *Angry* embeddings on top and considered the positions of other emotion embeddings across the x- and y-axis as separation across valence and arousal axes on the arousal-valence plane [3]. Notice better clustering by CQT embeddings at the emotion level. Nevertheless, both the embeddings show discrimination mainly across the arousal (vertical) axis. No major separation across the valence (horizontal) scale is observed across the embeddings. This indicates that the time-frequency representation can distinguish emotions across the arousal scale more than the valence scale.

6.2. Comparison of SER performances

Table 4 shows the results obtained for different features, while Fig. 13 provides the visual representation of the numerical values in the table. Both constant-Q filterbank representations (CQT and CWT) perform better than MFSC in predicting emotion classes across different databases. This demonstrates the inappropriateness of mel-scale for emotion prediction. The considerable improvement in constant-Q representations across different architectures justifies the salience of low-frequency regions of speech for SER. Between the constant-Q features, the performance varies across databases. CWT is either equivalent to or better than CQT over different databases. For EmoDB, CWT mostly shows higher UAR than CQT. For RAVDESS, CQT is better than CWT, whereas, in IEMOCAP and eNTERFACE, CWT performs better than CQT. Over RAVDESS database, an anomaly is the performance of CWT which is equivalent to MFSC with TDNN and Conv2D architectures.

Among the TDNN, Conv2D, and Conv2D-LSTM architectures, Conv2D-LSTM fairs better over every database. Between TDNN and Conv2D, the latter performs better in



(a) t-SNE plot of CQT embeddings

(b) t-SNE plot of MFSC embeddings

Fig. 12: t-SNE plots of CQT and MFSC embeddings for different speakers of EmoDB. The embeddings were extracted from the TDNN architecture. The points in every scatter plot are rotated to keep the *Anger* embeddings on top. The relative positioning of embeddings of other emotions provides an estimate of embedding's spread across the Arousal-Valence axis. Although no particular pattern is observed across the valence (horizontal) axis, embeddings form clusters across the arousal (vertical) scale. CQT embeddings show better clustering of emotions than MFSC.

Table 4: SER performance (in %) of CQT, MFSC, and CWT-based systems with different classifier back-ends. The **boldface** values show the highest performance metric obtained over different databases. Figure 13 provides a visual representation of the values in this table.

Database	**Architecture**	**Accuracy**			**UAR**		
		CQT	**MFSC**	**CWT**	**CQT**	**MFSC**	**CWT**
EmoDB	Conv2D	67.15	52.97	66.60	58.75	48.74	**62.24**
	TDNN	64.20	51.27	63.31	55.97	49.73	57.16
	Conv2D + LSTM + Attention (MERC 2020)	54.25	45.90	55.83	49.96	42.62	50.01
	Attention-based LSTM	61.44	42.83	59.08	57.57	40.92	53.97
	Transformer encoder model	48.89	46.16	51.21	42.32	41.39	45.52
	ECAPA-TDNN	54.71	48.30	52.93	47.21	44.98	45.89
	Conv2D-LSTM	65.69	55.51	**67.36**	60.45	50.08	61.76
RAVDESS	Conv2D	39.16	38.68	38.05	36.96	35.70	35.64
	TDNN	40.34	35.00	35.00	36.74	30.89	31.23
	Conv2D + LSTM + Attention (MERC 2020)	37.51	32.40	37.56	35.24	31.60	35.21
	Attention-based LSTM	43.12	32.08	42.50	42.74	31.62	40.58
	Transformer encoder model	35.69	29.93	34.16	33.23	27.96	32.28
	ECAPA-TDNN	32.43	34.37	34.86	31.82	33.61	34.16
	Conv2D-LSTM	**46.94**	42.56	46.43	44.15	39.80	**44.60**
eNTERFACE	Conv2D	46.77	41.86	49.09	41.31	37.77	43.35
	TDNN	52.65	35.72	56.91	39.27	33.05	41.92
	Conv2D + LSTM + Attention (MERC 2020)	45.33	44.58	45.49	43.35	43.07	43.82
	Attention-based LSTM	41.10	40.02	42.92	40.43	39.16	42.61
	Transformer encoder model	38.87	41.55	44.78	37.96	38.69	42.01
	ECAPA-TDNN	55.85	52.73	54.73	**55.10**	52.08	54.50
	Conv2D-LSTM	48.36	47.39	**59.05**	46.76	44.84	54.80
IEMOCAP	Conv2D	53.08	48.77	55.38	45.04	40.20	45.85
	TDNN	53.45	52.36	55.25	42.07	40.14	42.40
	Conv2D + LSTM + Attention (MERC 2020)	50.16	47.16	50.67	43.42	39.78	44.14
	Attention-based LSTM	47.41	45.68	51.50	44.22	40.41	46.73
	Transformer encoder model	48.77	43.83	51.45	41.40	38.73	42.53
	ECAPA-TDNN	50.32	46.19	48.33	42.77	40.90	42.53
	Conv2D-LSTM	54.83	48.34	**57.31**	46.53	41.11	**47.94**

terms of UAR except in EmoDB with the MFSC feature. This shows a better capability of convolution layers in improving prediction accuracy across all emotions instead of focusing on the more dominant emotion classes. The 2-dimensional convolution in Conv2D utilizes the correlation of features across both time and frequency domains and extracts variations across smaller time-frequency windows defined by the filter size at every layer. This helps accumulate refined emotion information from filters of varying sizes across

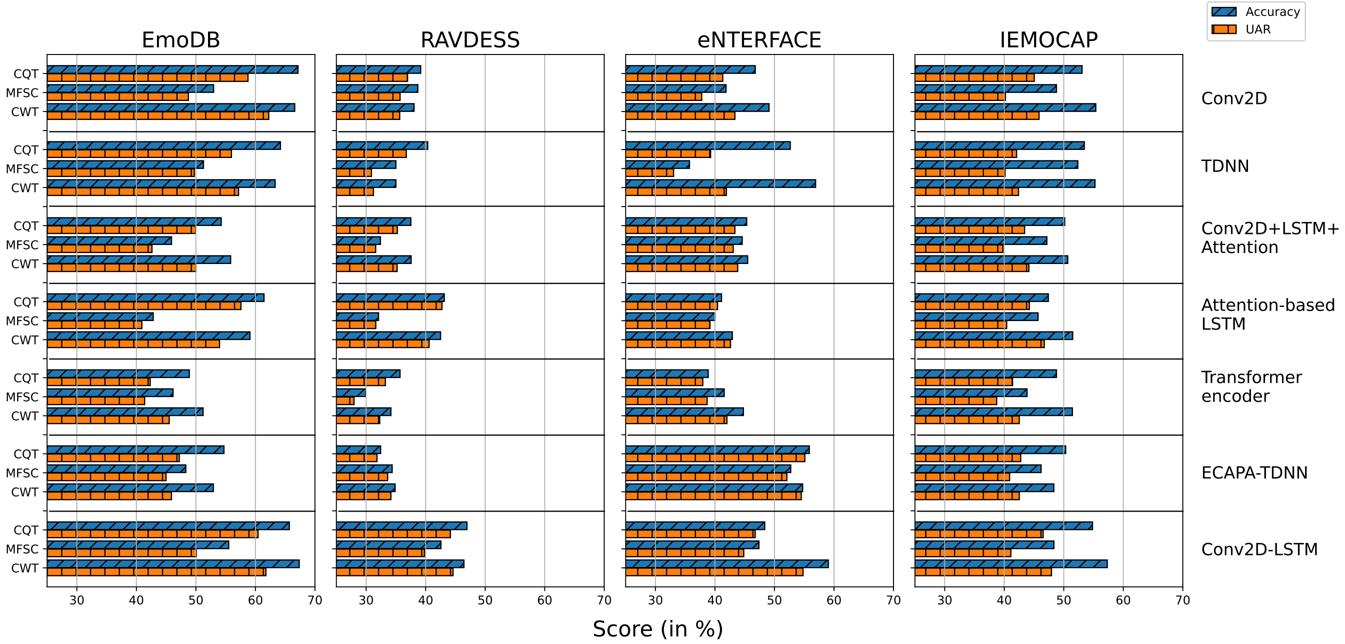


Fig. 13: Bar graph showing accuracy and UAR obtained with different features, neural network architectures, and databases. This is the visual representation of Table 4.

different layers of the model, which generates a better representation of emotion at the output. The LSTM layer further extracts the required temporal information from the activation of Conv2D layers, thereby improving performance.

For every architecture, constant-Q features outperform MFSC over every database. Also, the selected state-of-the-art deep learning architectures, i.e., Attention-based LSTM, Transformer encoder model, and ECAPA-TDNN, show inferior performance compared to plain Conv2D, TDNN, and Conv2D-LSTM. The Transformer encoder model has the poorest results among state-of-the-art networks across different databases because of the presence of only non-linear transformation-based attention layers and the lack of filter-based convolution operation to extract both time and frequency-based features. However, Attention-based LSTM architecture provides comparable performance (at least in UAR) to plain Conv2D and TDNN architectures. The final time and frequency-based attention layers in Attention-based LSTM help focus more on the emotion-relevant time and frequency cues. Similarly, ECAPA-TDNN outperforms both plain Conv2D and TDNN on eNTERFACE (in UAR) and provides comparable results on IEMOCAP. As ECAPA-TDNN contains many network parameters, its higher performance on databases with more data samples is justified. The Conv2D-LSTM model combines convolutional and temporal information extraction (due to the recurrent layer) to provide better performance on most databases.

Figure 14 shows the comparison of emotion-wise accuracy obtained with different features over plain Conv2D architecture. The performance of constant-Q features is similar across different emotions, except for *Neutral* and *Disgust*. As the CWT includes fram-

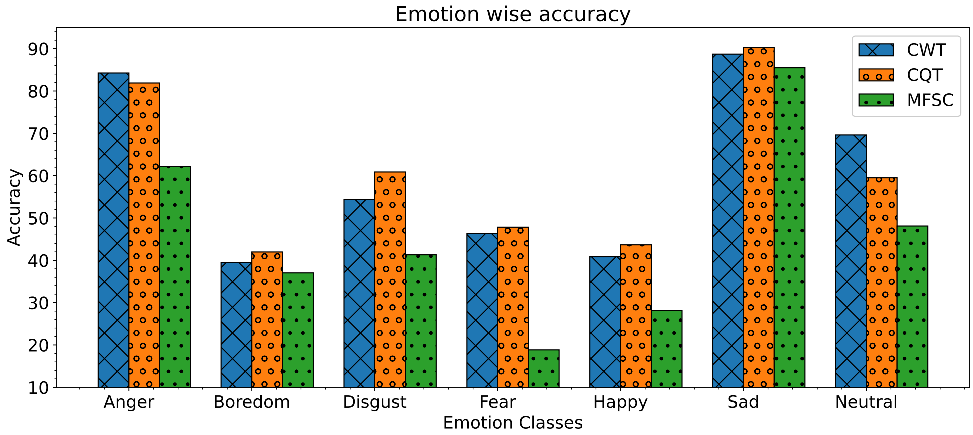


Fig. 14: Emotion-wise performance comparison among CQT, CWT, and MFSC on EmoDB database. The accuracies are computed with Conv2D architecture as classifier. Constant-Q features provide greatest advantage over MFSC on high arousal emotion (e.g., *Fear, Anger*).

ing and windowing, the emotions less sensitive to high-frequency content and benefited from high-frequency averaging are emphasized by CWT. Figure 14 shows that *Anger*, which is known to have greater high frequency relevance [34, 35], gains slightly from time-invariance (or averaging) applied at high frequency. The same is observed for *Neutral*. In contrast, *Disgust*, which is also sensitive to high-frequency variations ([35]), is represented better with CQT rather than with averaged high-frequency representation in CWT.

The improvement with constant-Q features is more than that for MFSC, especially on high-arousal emotions (*Fear, Anger, Happy*) because of the higher pitch resolution in the former. Contours of pitch harmonics in high-arousal emotions contain sudden rises and drops. These intonations are better captured in constant-Q features with higher resolution at low frequencies. For low-arousal emotions like *Sad* and *Boredom*, pitch contours usually follow straight-line patterns causing similar performance of MFSC and constant-Q features. For every feature, *Sad* emotion class yields the highest classification accuracy. MFSC provided the lowest accuracy score for *Fear* whereas the constant-Q representations yield the lowest accuracy score for *Boredom*. The difference between recognition accuracies of *Sad* and *Boredom* appears mainly because of similar arousal and valence characteristics.

Table 5 lists a few recent and relevant SER works. The absence of consensus on a standard experimental setup in SER literature is evident in the table. High SER performance can be achieved with less stringent evaluation frameworks, e.g., speaker-dependent train/test split, fewer cross-validation folds, fewer emotion classes from the database, etc. Thus, the performance comparison of multiple systems is difficult and mostly non-conclusive. Other factors like the absence of standard performance metric and lack of reproducible research in SER adds to inaccuracy in comparison. In our attempt to compare the employed system with other SER works, we include both methodology and experimental setup in the table so that the differences among works are understood

Table 5: Performance comparison with relevant works. Abbreviations: eGeMAPS = extended Geneva minimalistic acoustic parameter set, SVM = Support vector machine, COVAREP = Collaborative voice analysis repository, ECAPA-TDNN = Emphasized channel attention, propagation and aggregation in TDNN, MFLE = Mel-frequency log energy, CLDNN = Convolutional bidirectional LSTM DNN, LOSGO = Leave one speaker group out.

Reference	**Methodology**	**Evaluation method**	**Acc./UAR** (in %)
EmoDB			
Parry et al. (2019)	MFSC feature, CNN-LSTM classifier	80:10:10% train/valid/test split, 7 emotion classes	-/69.72
Triantafyllopoulos et al. (2019)	eGeMAPS, ComParE features, SVM classifier	LOSO cross-validation, 7 emotion classes	eGeMAPS: -/61.04 ComParE: -/72.34
Haider et al. (2021)	eGeMAPS features, No feature selection, SVM classifier	LOSO cross-validation, 7 emotion classes	-/68.5
Ours (Best result)	CWT feature, Conv2D-LSTM classifier	LOSO cross-validation, 7 emotion classes	67.36/61.76
RAVDESS			
Guizzo et al. (2020)	Spectrogram applied to multi-time-scale learning with CNN classifier.	4-fold cross-validation, 70:20:10% train:valid:test split, 8 emotion classes	55.85/-
Dissanayake et al. (2020)	MFCC feature, CNN with autoencoder and LSTM classifier	Training:Valid:Test; 22:1:1 speakers, 8 classes mapped to 3 classes, Positive, Negative, and Neutral	-/56.71
Beard et al. (2018)	COVAREP features, LSTM classifier	Evaluation method not defined, 8 emotion classes	41.25/-
Ours (Best result)	CWT feature, Conv2D-LSTM classifier	LOSO cross-validation, 8 emotion classes	46.43/44.60
eNTERFACE			
Triantafyllopoulos et al. (2019)	eGeMAPS, ComParE features, SVM classifier	LOSO cross-validation, 6 emotion classes	eGeMAPS:-/47.87 ComParE:-/65.60
Zhang et al. (2017)	MFSC feature (no delta, double-delta), DCNN (AlexNet) with average pooling, SVM classifier	LOSGO cross-validation, 6 emotion classes	51.33/-
Ours (Best result)	CWT feature, Conv2D-LSTM classifier	LOSO cross-validation, 6 emotion classes	59.05/54.80
IEMOCAP			
Pandey et al. (2022)	MFSC feature, CNN-LSTM as classifier	LOSO cross-validation, 4 emotion classes	51.82/-
Kumawat and Routray (2021)	MFCC feature, ECAPA-TDNN classifier	Leave-one-session-out cross validation, 4 emotion classes	58.76/-
Dissanayake et al. 2020	MFCC feature, CNN with autoencoder and LSTM classifier	5 sessions in training, half of 6th session in valid, half in test, 4 emotion classes	-/46.79
Meyer et al. (2018)	MFLE feature, CLDNN classifier	LOSGO cross-validation, 4 emotion classes	-/59.5
Ours (Best result)	CWT feature, Conv2D-LSTM classifier	LOSO cross-validation, 4 emotion classes	57.31/47.94

and, to some extent, the comparison is valid.

Comparison with the referred works shows that our method outperforms most of them on eNTERFACE and IEMOCAP databases but not EmoDB and RAVDESS. In EmoDB, the less stringent evaluation protocol in [20] and [80], in terms of train/test split, provides better classification results. The same argument applies to the works on the RAVDESS database. Although higher results can be achieved with lenient evaluation strategies, they do not imitate the real-world SER testing scenario. As discussed in Section 5.3, the LOSO cross-validation extensively evaluates the generalizability of a system with input from different unseen speakers without a large increase in computation complexity with small databases, making it a better evaluation strategy, especially for SER.

Let us now consider the works that use large feature sets [79, 80]. The famous eGeMAPS [10] and ComParE [87] feature sets contain many handcrafted features, including a combination of spectral and prosody features with statistics of a different order, unlike our method, which used spectral information only. Although considered appropriate as baseline, eGeMAPS and ComParE feature sets were designed after reviewing and selecting various handcrafted features found successful in previous SER studies [10]. Therefore, using these feature sets can also be considered human intervention in the train/test phases of the machine learning system contrary to using a specific handcrafted time-frequency representation for supervised learning and hence, also explains the performance difference from our method.

6.3. Time-invariance analysis over features

The increased low-frequency resolution in constant-Q filterbank representation brings out better resolution of pitch information (Section 4.4: Fig. 10) thereby enhancing emotion recognition performance (Section 6.2: Table 5, Fig. 13, and Fig. 14). Regarding the difference in time-invariant property between MFSC and constant-Q-based responses, we compared the temporal lengths of optimized CQT basis and mel-filters in Fig. 15. The low-frequency CQT basis functions (bases with center frequency $\leq$ 262 Hz) have a temporal spread greater than that of the 20 ms window used in STFT. Also, the modulus operation on computed CQT coefficients removes the phase information, leading to time-invariance defined by the length of the basis functions. Human pitch frequencies exist at around 250 Hz [88]. Therefore, the time-invariance in CQT is greater than that of the standard STFT around pitch frequencies, making the former more robust against emotion-irrelevant pitch variations. Notice that the temporal spread of mel-filters remains below 20 ms over the complete frequency range. Hence, in MFSC, the temporal invariance is defined by the window length (ϕ) fixed to 20 ms. In CWT, we applied a similar averaging of computed CWT coefficients by a 20 ms window causing a varying time-invariance when the CWT basis is greater than the window length and fixed time-invariance otherwise. The results show that the combination of varying and fixed time-invariance leads to slight improvement across most emotion-database pairs. Additionally, increased time-resolution at high frequencies in constant-Q filterbank response contributes less to SER. Rather, capturing long temporal information is a better approach. The long time-scale information alongside the translation invariant feature learning capability of deep networks (obtained from convolutional layers, statistics pooling layer, etc.) improved SER performance in our experiments.

Although our CWT implementation is very similar to the first layer scattering coefficients [62] (Section 4), unlike the former, the scattering coefficients provide fixed

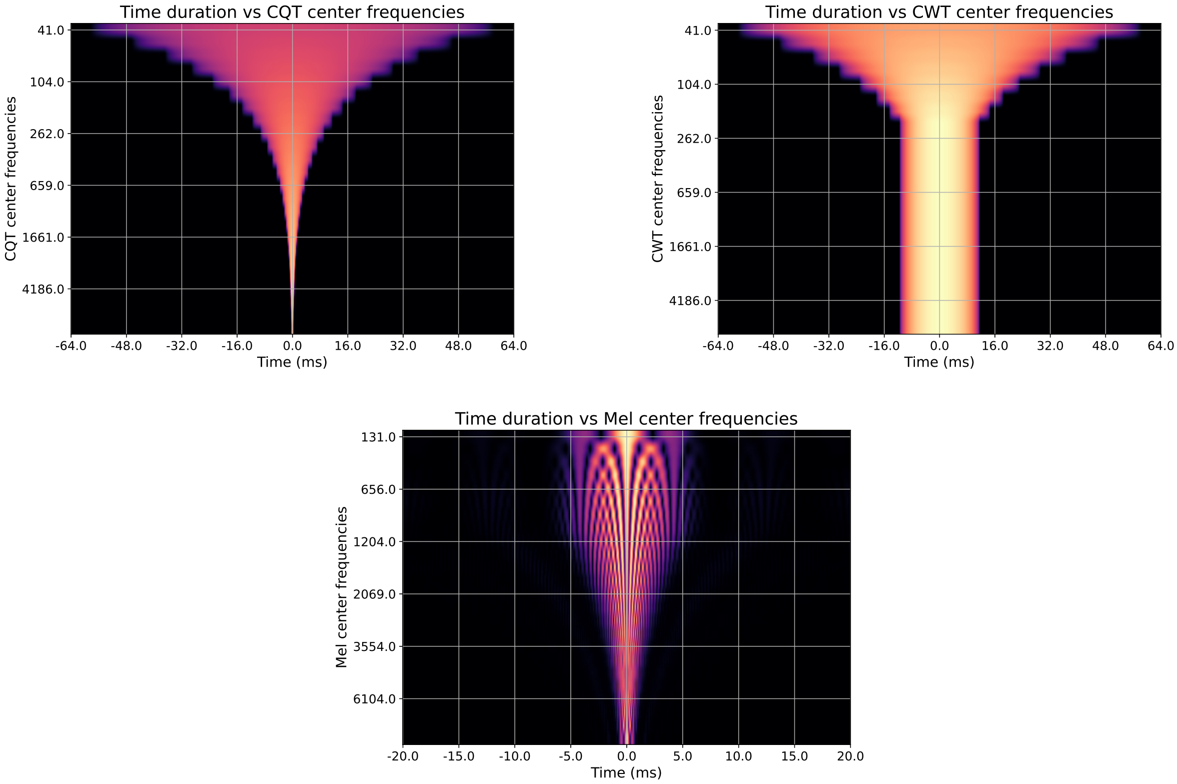


Fig. 15: Time invariance provided by CQT, CWT, and MFSC with changing frequency. y-axis: the center frequency of filters (or frequency bins), 24 filters per feature; x-axis: the filters' corresponding time spread. Both CQT and CWT use basis functions with similar time spread. However, framing in CWT makes time invariance fixed to 20 ms at mid and high-frequency bins. Mel-filters have a much lower time spread than constant-Q filters at low frequencies.

time-invariance over the complete frequency range. In a separate study, we found that our CWT implementation also performs better than the first layer scattering coefficients for SER [89].

6.4. Complexity analysis

To compare the complexity of different features, we calculated the floating-point operations (FLOPs) and the time required to compute the features. Table 6 reports features' FLOPs and computation time averaged across 100 runs of feature extraction of a randomly chosen one-second speech sample. The FLOPs count estimates the number of mathematical operations (additions, subtractions, multiplications, and divisions) required to compute the feature. We calculated FLOPs during feature extraction using only one CPU core (Intel Xeon E5-2670 2.6 GHz) and Linux `perf`[10] command. The

[10] `https://perf.wiki.kernel.org/index.php/Main_Page`

Table 6: FLOPs and time required for computation of different features. MFLOPs = Mega FLOPs.

Complexity Parameter	MFSC	CQT	CWT
FLOPs	8.60 MFLOPs	6.73 MFLOPs	110.88 MFLOPs
Time	1.047 s	1.629 s	1.118 s

table shows that the number of FLOPs and computation time is inconsistent across features. For example, CQT requires fewer FLOPs but more time to compute, while CWT requires more FLOPs but less computation time. This inconsistency might result from how processors compute integers, fractions, etc., and the time it takes to perform such computations. The CQT requires fewer floating-point operations, but they might take more time to compute, whereas CWT needs more operations but might take less time to calculate.

7. Conclusion

This paper presented the role of constant-Q filterbank based time-frequency representations, namely CQT and CWT, for SER and compared them with the traditional mel-scaled representation. Our analysis expounded the emotion-relevant advantages provided by the former representations in both time and frequency domains. The comparison with the latter also showed that the greater emphasis provided by the constant-Q non-linearity over low-frequency regions accounts for better suitability of constant-Q representations for SER. The superiority remained consistent across different neural network architectures. From the experiments and analysis, we conclude the following:

1. Constant-Q filterbank representation provides higher time-invariance and increased frequency resolution at low speech-frequencies, causing an improved SER performance compared to mel-scale based representation.
2. In the time-domain, CQT provides time-invariance increasing towards low frequencies leading to robustness against emotion irrelevant temporal variations and eventually better emotion prediction.
3. The CWT implementation bears a combination of varying and fixed time-invariance over different frequency bins and offers an advantage similar to CQT in SER performance. The difference in performance between the two is attributed to the fixed time-invariance in CWT at mid and high frequencies.
4. Like mel-scale representation, constant-Q representations also provide stability to time-warp deformations begetting a robust descriptor in terms of variations due to different speaking styles.
5. In the frequency domain, constant-Q filterbank representations are more efficient in resolving pitch harmonics than mel-scaled representations contributing to a better SER performance because of the higher relevance of pitch in emotion prediction.
6. Constant-Q representations outperform mel-scale representations over multiple neural network architectures.

7. Studies in psychology show better emotion perception abilities of musicians than non-musicians [58, 59]. As CQT is a more appropriate representation for music analysis [43], better SER capabilities of CQT over MFSC suggest some linkage between human emotion and music perception.

Although there was a noticeable improvement with constant-Q filterbank representations, there is a need for further effort to develop a deployable real-world SER system. Table 6 shows approximately 60% and 10% more computation time for CQT and CWT features, respectively, compared with MFSC. This increase is the cost for better performance calling for further exploration of time-frequency representations for SER tasks. Future directions of this work include cross-corpora evaluation to study the generalization ability of constant-Q representation with out-of-domain training data. The SER datasets are small-scale datasets that limit the investigations of large-scale deep architectures for this task. The constant-Q representation can be examined in a transfer learning framework which involves learning a pretext task first on a large dataset followed by training downstream emotion recognition task on the limited size dataset.

References

[1] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020).

[2] S. R. Krothapalli, S. G. Koolagudi, Speech emotion recognition: A review, in: Emotion Recognition using Speech Features, Springer New York, New York, NY, 2013, pp. 15–34.

[3] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (2011) 572–587.

[4] R. W. Picard, Affective Computing, MIT Press, 2000.

[5] R. W. Picard, Affective computing: challenges, International Journal of Human-Computer Studies 59 (2003) 55–64.

[6] A. H. Fischer, P. M. Rodriguez Mosquera, A. E. Van Vianen, A. S. Manstead, Gender and culture differences in emotion, Emotion 4 (2004) 87.

[7] G. Bryant, H. C. Barrett, Vocal emotion recognition across disparate cultures, Journal of Cognition and Culture 8 (2008) 135–148.

[8] N. Lim, Cultural differences in emotion: Differences in emotional arousal level between the East and the West, Integrative Medicine Research 5 (2016) 105–109.

[9] F. Eyben, A. Batliner, B. Schuller, Towards a standard set of acoustic features for the processing of emotion in speech, in: Proc. Meetings on Acoustics, volume 9, 2010, p. 060006.

[10] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Transactions on Affective Computing 7 (2016) 190–202.

[11] L. Chen, X. Mao, Y. Xue, L. L. Cheng, Speech emotion recognition: Features and classification models, Digital Signal Processing 22 (2012) 1154–1160.

[12] Y. Zhou, Y. Sun, J. Zhang, Y. Yan, Speech emotion recognition using both spectral and prosodic features, in: Proc. International Conference on Information Engineering and Computer Science, 2009, pp. 1–4.

[13] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, et al., Whodunnit–searching for the most important feature types signalling emotion-related user states in speech, Computer Speech & Language 25 (2011) 4–28.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: Proc. ICASSP, 2016, pp. 5200–5204.

[15] C.-W. Huang, S. Narayanan, Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition, arXiv preprint arXiv:1706.02901, 2017.

[16] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, IEEE Transactions on Multimedia 16 (2014) 2203–2213.

[17] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Transactions on Multimedia 20 (2017) 1576–1590.

[18] Z. C. Lipton, The mythos of model interpretability, Queue 16 (2018) 31–57.

[19] D. Rolnick, A. Veit, S. Belongie, N. Shavit, Deep learning is robust to massive label noise, arXiv preprint arXiv:1705.10694, 2017.

[20] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, G. Hofer, Analysis of deep learning architectures for cross-corpus speech emotion recognition, in: Proc. INTERSPEECH, 2019, pp. 1656–1660.

[21] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, in: Proc. ICSLP, volume 3, 1996, pp. 1970–1973.

[22] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, S. Stroeve, Approaching automatic recognition of emotion from voice: A rough benchmark, in: Proc. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, 2000.

[23] S. E. Bou-Ghazale, J. H. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Transactions on Speech and Audio Processing 8 (2000) 429–442.

[24] T. L. Nwe, S. W. Foo, L. C. De Silva, Speech emotion recognition using hidden Markov models, Speech Communication 41 (2003) 603–623.

[25] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech Communication 52 (2010) 613–625.

[26] S. Wu, T. H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Communication 53 (2011) 768–785.

[27] K. Wang, N. An, B. N. Li, Y. Zhang, L. Li, Speech emotion recognition using Fourier parameters, IEEE Transactions on Affective Computing 6 (2015) 69–75.

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

[29] D. Tang, J. Zeng, M. Li, An end-to-end deep learning framework for speech emotion recognition of atypical individuals, in: Proc. INTERSPEECH, 2018, pp. 162–166.

[30] S. Ghosh, E. Laksana, L.-P. Morency, S. Scherer, Representation learning for speech emotion recognition, in: Proc. INTERSPEECH, 2016, pp. 3603–3607.

[31] H. M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Networks 92 (2017) 60–68.

[32] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Language Resources and Evaluation 42 (2008) 335–359.

[33] D. Mika, J. Józwik, Advanced time-frequency representation in voice signal analysis, Advances in Science and Technology Research Journal 12 (2018) 251–259.

[34] C. E. Williams, K. N. Stevens, Emotions and speech: Some acoustical correlates, The Journal of the Acoustical Society of America 52 (1972) 1238–1250.

[35] R. Banse, K. R. Scherer, Acoustic profiles in vocal emotion expression, Journal of Personality and Social Psychology 70 (1996) 614.

[36] R. Cowie, E. Douglas-Cowie, Automatic statistical analysis of the signal and prosodic signs of emotion in speech, in: Proc. ICSLP, volume 3, 1996, pp. 1989–1992.

[37] S. Deb, S. Dandapat, Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification, IEEE Transactions on Cybernetics 49 (2018) 802–815.

[38] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE Transactions on Biomedical Engineering 47 (2000) 829–837.

[39] M. Goudbeek, J. P. Goldman, K. R. Scherer, Emotion dimensions and formant position, in: Proc. INTERPSEECH, 2009, pp. 1575–1578.

[40] E. Bozkurt, E. Erzin, C. E. Erdem, A. T. Erdem, Formant position based weighted spectral features for emotion recognition, Speech Communication 53 (2011) 1186–1197.

[41] M. Lech, M. Stolar, R. Bolia, M. Skinner, Amplitude-frequency analysis of emotional speech using transfer learning and classification of spectrogram images, Advances in Science, Technology and Engineering Systems Journal 3 (2018) 363–371.

[42] H. M. Chandrashekar, V. Karjigi, N. Sreedevi, Investigation of different time-frequency representations for intelligibility assessment of dysarthric speech, IEEE Transactions on Neural Systems and Rehabilitation Engineering 28 (2020) 2880–2889.

[43] J. C. Brown, Calculation of a constant Q spectral transform, The Journal of the Acoustical Society of America 89 (1991) 425–434.

[44] M. Todisco, H. Delgado, N. Evans, Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification, Computer Speech & Language 45 (2017) 516–535.

[45] M. Pal, D. Paul, G. Saha, Synthetic speech detection using fundamental frequency variation and spectral features, Computer Speech & Language 48 (2018) 31–50.

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

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

[48] S. Waldekar, G. Saha, Classification of audio scenes with novel features in a fused system framework, Digital Signal Processing 75 (2018) 71–82.

[49] O. Rioul, M. Vetterli, Wavelets and signal processing, IEEE Signal Processing Magazine 8 (1991) 14–38.

[50] Y. Huang, A. Wu, G. Zhang, Y. Li, Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition, IET Signal Processing 9 (2015) 341–348.

[51] S. Ntalampiras, N. Fakotakis, Modeling the temporal evolution of acoustic parameters for speech emotion recognition, IEEE Transactions on Affective Computing 3 (2012) 116–125.

[52] K.-C. Wang, Time-frequency feature representation using multi-resolution texture analysis and acoustic activity detector for real-life speech emotion recognition, Sensors 15 (2015) 1458–1478.

[53] P. Shegokar, P. Sircar, Continuous wavelet transform based speech emotion recognition, in: Proc. ICSPCS, 2016, pp. 1–8.

[54] P. Singh, G. Saha, 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), 2021.

[55] C. Schörkhuber, A. Klapuri, Constant-Q transform toolbox for music processing, in: Proc. 7th Sound and Music Computing Conference, Barcelona, Spain, 2010, pp. 3–64.

[56] J. Yang, R. K. Das, Improving anti-spoofing with octave spectrum and short-term spectral statistics information, Applied Acoustics 157 (2020) 107017.

[57] S. Nicholson, B. Milner, S. Cox, Evaluating feature set performance using the F-ratio and J-measures, in: Proc. EUROSPEECH, 1997, pp. 413–416.

[58] C. F. Lima, S. L. Castro, Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody, Emotion 11 (2011) 1021.

[59] A. Good, K. A. Gordon, B. C. Papsin, G. Nespoli, T. Hopyan, I. Peretz, F. A. Russo, Benefits of music training for perception of emotional speech prosody in deaf children with cochlear implants, Ear Hear 38 (2017) 455.

[60] O. Rioul, M. Vetterli, Wavelets and signal processing, IEEE Signal Processing Magazine 8 (1991) 14–38.

[61] M. Sahidullah, G. Saha, Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition, Speech Communication 54 (2012) 543–565.

[62] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Transactions on Signal Processing 62 (2014) 4114–4128.

[63] X. Lu, J. Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification, Speech Communication 50 (2008) 312–322.

[64] S. Sarangi, M. Sahidullah, G. Saha, Optimization of data-driven filterbank for automatic speaker verification, Digital Signal Processing 104 (2020) 102795.

[65] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, B. Schuller, Speech emotion classification using attention-based LSTM, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (2019) 1675–1685.

[66] Z. Lian, B. Liu, J. Tao, CTNet: Conversational transformer network for emotion recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021) 985–1000.

[67] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, H. Na, ECAPA-TDNN embeddings for speaker diarization, in: Proc. INTERSPEECH, 2021, pp. 3560–3564.

[68] B. Desplanques, J. Thienpondt, K. Demuynck, ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, 2020, pp. 3830–3834.

[69] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: Proc. ICASSP, 2018, pp. 5329–5333.

[70] M. Chen, X. He, J. Yang, H. Zhang, 3-D convolutional recurrent neural networks with attention model for speech emotion recognition, IEEE Signal Processing Letters 25 (2018) 1440–1444.

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

[72] S. R. Livingstone, 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 13 (2018).

[73] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE'05 audio-visual emotion database, in: Proc. International Conference on Data Engineering Workshops, 2006, pp. 8–8.

[74] D. Issa, M. F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control 59 (2020) 101894.

[75] D. Luo, Y. Zou, D. Huang, Investigation on joint representation learning for robust feature extraction in speech emotion recognition, in: Proc. INTERSPEECH, 2018, pp. 152–156.

[76] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, G. Rigoll, Speaker independent speech emotion recognition by ensemble classification, in: Proc. International Conference on Multimedia and Expo, 2005, pp. 864–867.

[77] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: A benchmark comparison of performances, in: Proc. Workshop on Automatic Speech Recognition Understanding, 2009, pp. 552–557.

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

[79] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, B. W. Schuller, Towards robust speech emotion recognition using deep residual networks for speech enhancement, in: Proc. INTERSPEECH, 2019, pp. 1691–1695.

[80] F. Haider, S. Pollak, P. Albert, S. Luz, Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods, Computer Speech & Language 65 (2021) 101119.

[81] E. Guizzo, T. Weyde, J. B. Leveson, Multi-time-scale convolution for emotion recognition from speech audio signals, in: Proc. ICASSP, 2020, pp. 6489–6493.

[82] V. Dissanayake, H. Zhang, M. Billinghurst, S. Nanayakkara, Speech emotion recognition 'in the wild' using an autoencoder, in: Proc. INTERSPEECH, 2020, pp. 526–530.

[83] R. Beard, R. Das, R. W. M. Ng, P. G. K. Gopalakrishnan, L. Eerens, P. Swietojanski, O. Miksik, Multi-modal sequence fusion via recursive attention for emotion recognition, in: Proc. Conference on Computational Natural Language Learning, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 251–259.

[84] S. K. Pandey, H. S. Shekhawat, S. Prasanna, Attention gated tensor neural network architectures for speech emotion recognition, Biomedical Signal Processing and Control 71 (2022) 103173.

[85] P. Kumawat, A. Routray, Applying TDNN architectures for analyzing duration dependencies on speech emotion recognition, in: Proc. INTERSPEECH 2021, 2021, pp. 3410–3414.

[86] P. Meyer, E. Buschermöhle, T. Fingscheidt, What do classifiers actually learn? a case study on emotion recognition datasets, in: Proc. INTERSPEECH, 2018, pp. 262–266.

[87] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini, et al., The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language, in: Proc. INTERPSEECH, 2016, pp. 2001–2005.

[88] J. M. Hillenbrand, M. J. Clark, The role of F0 and formant frequencies in distinguishing the voices of men and women, Attention, Perception, & Psychophysics 71 (2009) 1150–1166.

[89] P. Singh, G. Saha, M. Sahidullah, Deep scattering network for speech emotion recognition, in: Proc. EUSIPCO, 2021, pp. 131–135.