Modulation spectral features for speech emotion recognition using deep neural networks

Premjeet Singh, Md Sahidullah, Goutam Saha

Modulation spectral features for speech emotion recognition using deep neural networks

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

[a] *Department of Electronics & Electrical Communication Engineering Indian Institute of Technology, India-721302, Kharagpur, India*
[b] *Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France*

Abstract

This work explores the use of constant-Q transform based modulation spectral features (CQT-MSF) for speech emotion recognition (SER). The human perception and analysis of sound comprise of two important cognitive parts: early auditory analysis and cortex-based processing. The early auditory analysis considers spectrogram-based representation whereas cortex-based analysis includes extraction of temporal modulations from the spectrogram. This temporal modulation representation of spectrogram is called modulation spectral feature (MSF). As the constant-Q transform (CQT) provides higher resolution at emotion salient low-frequency regions of speech, we find that CQT-based spectrogram, together with its temporal modulations, provides a representation enriched with emotion-specific information. We argue that CQT-MSF when used with a 2-dimensional convolutional network can provide a time-shift invariant and deformation insensitive representation for SER. Our results show that CQT-MSF outperforms standard mel-scale based spectrogram and its modulation features on two popular SER databases, Berlin EmoDB and RAVDESS. We also show that our proposed feature outperforms the shift and deformation invariant scattering transform coefficients, hence, showing the importance of joint hand-crafted and self-learned feature extraction instead of reliance on complete hand-crafted features. Finally, we perform Grad-CAM analysis to visually inspect the contribution of constant-Q modulation features over SER.

Keywords: Constant-Q transform, Convolutional neural network, Modulation spectrogram, Gammatone spectrogram, Shift invariance, Speech emotion recognition.

1. Introduction

Speech emotion recognition (SER) is the process of automatic prediction of speaker's emotional state from his/her speech samples. A speech sample generally remains enriched with various information, such as speaker, language, emotion, context, recording environment, gender and age, intricately entangled to each other [1]. Human mind

*Corresponding author
Email addresses: premsingh@iitkgp.ac.in (Premjeet Singh), md.sahidullah@inria.fr (Md Sahidullah), gsaha@ece.iitkgp.ernet.in (Goutam Saha)

is congenitally trained to disentangle such information, however, same is not true for machines [2]. Machines need to be specifically trained to extract cues pertaining to a particular information. Among such, extraction of emotion-specific cues for SER is still considered a challenging task. The challenge basically persists because of the differences in the manner of emotion expression across individuals [3]. These differences stem from factors such as speaker's culture and background, ethnicity, speaker's mood, gender, manner of speech, etc. [3, 4]. For automatic SER, a machine should be capable of extracting emotion-specific cues in the presence of all such variabilities.

SER finds application in several human-computer interaction domains such as sentiment analysis in customer service, health care systems, self-driving vehicles, auto-pilot systems, product advertisement and analysis [1, 3, 5]. One of the first seminal works in SER was aimed towards emotion information extraction using different speech cues [6]. Various works that followed discovered that *speech prosody* (pitch, intonation, energy, loudness, etc.) contain significant information for emotion discrimination [7–9]. Similarly, several other works report that *spectral features* (spectral flux, centroid, mel-frequency cepstral coefficients (MFCCs), etc.) and *voice-quality features* (jitter, shimmer, harmonic-to-noise ratio (HNR), etc.) of speech are also important for SER [10]. For classification, these extracted features are processed with a classifier back-end such as *support vector machine* (SVM), *Gaussian mixture model* (GMM), and *k-nearest neighbour* (k-NN) for emotion class prediction. These approaches which employ certain signal processing algorithm for feature extraction are termed hand-crafted feature based approaches for SER. Hand-crafted approaches enjoy the advantage of being interpretable, in terms of which feature or speech characteristic is more relevant for emotions, and are computationally inexpensive. However, hand-crafted features often suffer from *curse of dimensionality*, especially when *brute-force* method based SER system is used [11].

Recent advancements in signal processing have introduced *deep neural networks* (DNN) into the speech processing domain. DNNs have the impressive ability by which, given the required data, they automatically learn to obtain a possible solution to pattern recognition problems. This is accomplished by automatically updating the DNN parameters so as to reduce the defined loss function and approach towards the local minima. In SER system, deep networks are either used as automatic feature extractors or as classifiers for emotion class prediction. Recently, a new deep learning paradigm is also introduced which performs both feature extraction and emotion classification in an end-to-end fashion. Along these lines, several works use *convolutional neural network* (CNN) as automatic feature extractor for SER [12, 13]. In contrast, other approaches use hand-crafted methods for feature extraction which are then used as features for DNN classifier input [14–16]. To obtain an end-to-end solution for SER works in [17–19] have used DNNs where the initial layers extract the emotion-relevant features and final layers act as classifier. In recent years, deep learning methods have been consistently shown to outperform hand-crafted feature based SER techniques.

In spite of their tremendous success, DNNs have major practical disadvantages. One such disadvantage is the requirement of large labelled database for proper DNN training [20]. In contrast to other speech classification problems, such as speech and speaker recognition, large speech corpora are not available for evaluating SER task. Various

ethical and legal issues make it difficult to collect large dataset of natural emotional voices from real-world scenario [3, 5]. To somewhat alleviate this issue, acted emotion recordings are generally used where skilled actors enact a predefined set of emotions. However, this approach is not considered very appropriate as acted emotions are often exaggerated versions of natural emotions [3, 5]. Another disadvantage of DNN is its complexity. Due to large trainable parameter set and high non-linear relationship between input and output, DNNs are often termed *black-box* models, which are very difficult to understand/interpret [21–23]. As the training includes optimization of all DNN parameters, it takes much larger time for DNNs to train as compared to classical statistical methods (e.g., SVM or GMM). Hence, even though very appealing, DNN models are still far-off a completely optimized SER approach.

The above discussion leads to the conclusion that both hand-crafted and DNN based feature extraction methods have their own set of advantages and disadvantages. In this work, we aim to exploit the advantages of both the methods for improved SER performance. Our framework is similar to the combination of hand-crafted feature, in the form of time-frequency representation, and a DNN model for further feature enrichment as used in other SER works [14–16]. However, our approach incorporates the prior (domain) knowledge of speech processing in humans, i.e., early auditory and cortex-based processing of speech [24], for an improved hand-crafted feature representation. Being data-driven, DNN-based machine learning approaches suffer in performance, especially when there are constraints over the training data, e.g., limited size, ethical concerns in recording and poor quality of data [25], all of which are relevant for SER databases. Evidences reveal that such disadvantages can be alleviated by the use of domain knowledge [25, 26] in hand-crafted feature generation. Further, regarding speech representations, spectrogram and mel-spectrogram are considered the *de-facto* standard of time-frequency representations in SER. However, they encompass only the early auditory processing of speech and lack cortical information.

Inspired by this fact, we first employ a hand-crafted feature extraction technique which combines an emotion relevant early auditory representation with corresponding cortex-based representation of the speech for SER. These features are then processed by a deep convolutional neural network which further extracts the emotion relevant information. Two machine learning frameworks are used at the back-end: Convolutional network with fully connected layer, and convolutional layer for embedding extraction with SVM classifier for final emotion class prediction. Such combination of multi-stage hand-crafted feature with DNN at back-end more closely follows the natural speech processing workflow in humans where the auditory system captures the signal and extracts the features, which are then transmitted to the inner regions of brain for further analysis and understanding. The achieved improvement in performance over different databases further consolidates our hypothesis of two-staged hand-crafted speech processing for SER. Figure 1 provides a general overview of the two-staged processing framework in human auditory system for SER.

In the next section (Section 2), we describe the relevant literature and discuss the motivation and major contributions of this work. Section 3 and 4 provides a brief introduction to the early auditory and cortex-based feature representations used in this

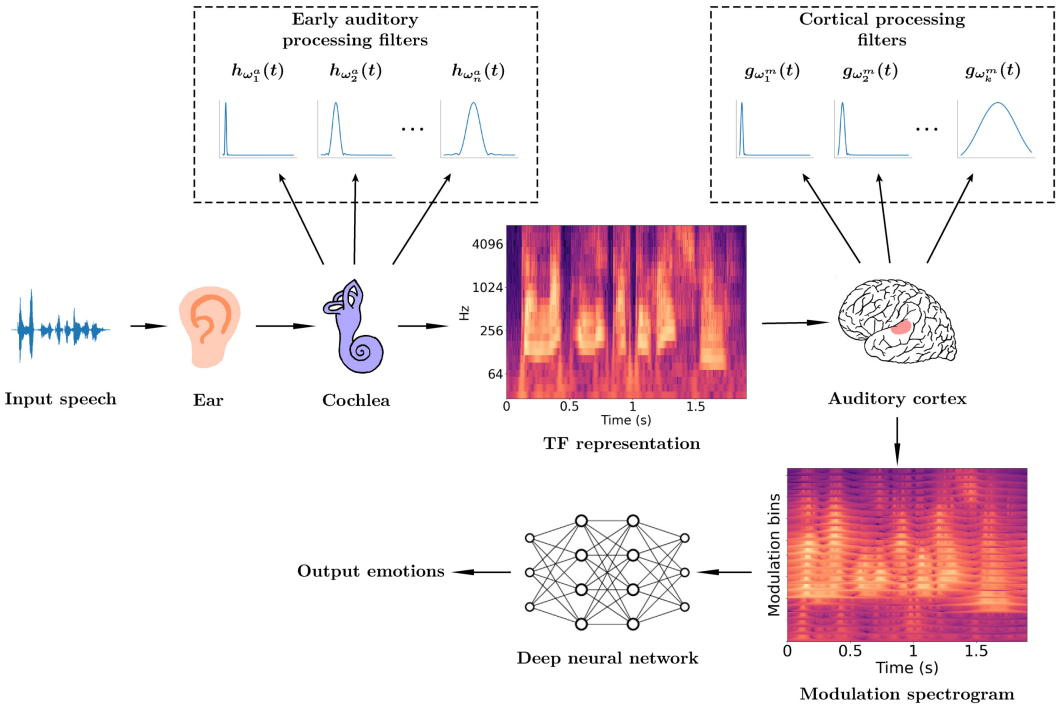


Fig. 1: The two-staged speech processing in the human auditory system for SER. The input speech captured at ear is converted to a form similar to time-frequency (TF) representation by the early auditory processing filters present in cochlea. This representation is then passed on to the auditory cortex in the brain for processing with cortical filters. The highlighted part in brain image identifies the auditory cortex region of the brain. The cortical filter processing leads to a modulation spectrogram based representation. This is further processed by the inner regions of the brain to finally decode emotions. Our employed deep neural network, which is loosely based on the studies of the brain and nervous system, models the inner processing of the brain to identify the emotion classes from the input modulation spectrogram feature. The $h_{\omega_n^a}(t)$ and $g_{\omega_k^m}(t)$ depict the impulse response of nth early auditory and mth cortical processing filter, respectively. The Figure shows logarithm applied TF and modulation spectrogram representation.

work. Section 5 describes the experimental setup used to perform the experiments. Section 6 describes the results obtained with the proposed feature and comparison with the standard features followed by corresponding discussion. Finally, Section 7 includes the conclusive statements of the work.

2. Related Works and Motivation

In this section, we provide a brief review of works related to the frequency localisation of emotions. We then discuss some works which describe the relevance of modulation spectrogram in speech processing. This is followed by description of the motivation of our proposed feature and the major contributions of this work.

2.1. Literature Review

Several studies, aimed towards analysing the importance of spectral frequencies, have reported the prominence of low frequencies in SER. Authors in [27] report the promi-

nence of first formant frequency (F1) for recognition of *Anger* and second formant (F2) for recognition of *Neutral*. Studies performed in [28] found that high arousal emotions, e.g, *Anger*, *Happy*, have higher average F1 value and lower F2 value. They also found that positive valence emotions (e.g., *Happy*, *Pride*, *Relief*) have higher average F2 value. Authors in [29] also report discrimination between idle and negative emotions using the temporal patterns of first two formant frequencies. In [30], authors show that non-linear frequency scales (e.g., equivalent rectangular bandwidth (ERB), mel, logarithmic) when applied for sub-band partitioning and energy computation over discrete Fourier transform based spectrogram, results in improved SER accuracy. Such studies hint toward the requirement of a non-linear frequency scale based time-frequency representation with higher emphasis on low-frequency regions of speech.

Regarding human sound perception, evidences in literature suggest that the process of auditory signal analysis can be modelled into two stages: (i) *Early auditory stage*, which models the incoming audio signal into a spectrogram based representation. (ii) *Cortical analysis stage*, which extracts the spectro-temporal modulation relationship among different audio cues from the auditory spectrogram [24, 31]. Such modelling strategy has been found effective in the analysis of both speech and music signals [24]. The spectral and temporal modulation features of speech spectrogram are also highly related to speech intelligibility, noise and reverberation effects [31]. In [32], authors report that the spectro-temporal representation of audio (non-speech) signals with positive\negative valence is different from that of neutral sounds. They also report that spectral frequency and temporal modulation frequency can represent the valence information of sounds. Authors in [33] compared the temporal modulations of human speech with scream voice and concluded that slow temporal variations (< 20 Hz) contain most linguistic (both prosodic and syllabic cues) information.

In speech analysis, temporal modulation features are called *modulation spectral features* (MSFs). Owing to their relatedness to speech intelligibility, MSFs have been extensively used in speech processing. Some works also successfully explored modulation features for speaker identification, verification and audio coding [34, 35]. Author in [36] provides a comprehensive description of the history of the use of modulation features in speech recognition.

Modulation features have also been explored for emotion recognition in speech. Authors in [37] used MSF for emotion recognition and provided a detailed explanation of its relevance for SER. In [38], authors used a smoothed nonlinear operator to obtain the amplitude modulated power spectrum of the gammatone filterbank generated spectrogram and showed improvement over standard MFCC for SER. Authors in [39] studied the relationship between human emotion perception and the MSFs of emotional speech and concluded on the suitability of modulation features for emotion recognition. Authors in [40] used 3-D convolutions and attention-based recurrent networks to combine auditory analysis and attention mechanisms for SER. This work also explains that the temporal modulations extracted from auditory analysis contain periodicity information important for emotion recognition. In [41], various feature pooling, such as *mean*, *standard deviation*, and *kurtosis*, on frame-level measure of MSF to be used for "in-the-wild" dimension-based SER (dimensional SER includes projection of speech onto three emotion

dimensions: *valence*, *arousal* and *dominance)*. The authors report improvement in results over frame-wise modulation spectral feature baseline for various noise and reverberated speech scenarios. Similar MSF measures when used with Bag-of-Audio-Words (BoAW) approach showed SER improvement against environmental noise in [42]. In [43], authors use modulation spectral features with convolutional neural networks to discriminate between stress-based speech and neutral speech. The authors show that the modulation spectral features when used with CNN with the time frames intact (without statistics pooling over time frames of MSF) gives better performance, especially over increased number of target emotion classes. In [44], authors show that joint spectro-temporal modulation representation outperforms standard MFCC in emotion classification of noisy speech. Recently, the authors in [45] have also used MSF over cochleagram features with a long short term memory (LSTM) based system for dimensional SER. The work explains that arousal information can be characterised by the amplitude envelope of speech signal, whereas valence information is characterised by the temporal dynamics of amplitude envelope. Since it is difficult to obtain such dynamics from low-level descriptor (LLD) features, auditory analysis based temporal modulation features can potentially represent the required temporal dynamics for SER.

2.2. Motivation and Contributions

The literature in SER reveals two important speech characteristics for emotion prediction: the importance of low frequencies, and the importance of temporal modulations of spectrogram. To address the importance of low-frequency information, we use constant-Q transform (CQT) based time-frequency representation for SER. CQT provides higher frequency resolution and increased time invariance at low frequencies thereby emphasizing the low-frequency regions of speech [46]. This helps in better resolution of emotion salient frequency regions of speech and improved SER performance [47]. CQT is also known to provide a representation with visible pitch frequency and well-separated pitch harmonics [48]. Because of high relevance of pitch information in emotion discrimination, this property of CQT makes it more suitable for SER over standard mel-based features.

To further enhance the CQT-based system while utilising the understanding of domain knowledge of human auditory-cortical physiology [31], we propose to use temporal modulation of CQT spectrogram representation for SER. Specifically, we use CQT spectrogram representation for auditory analysis and extract temporal modulations of CQT by again using constant-Q filters, for cortical analysis. In this way, we obtain the temporal modulation of emotion salient low-frequency regions which are emphasized by CQT. Studies show that such use of constant-Q modulation filterbank better approximates the cortical sound processing in humans [49, 50]. The constant-Q factor characteristic of modulation filters also lead to higher resolution at lower modulation frequencies, hence, providing an arrangement that helps in identifying any deviation from general (or *Neutral*) speech modulation rate (2-4 Hz) [51]. Our choice of constant-Q filters in both stages is also inspired from the study of early auditory and cortical stages of mammalian auditory cortex [31, 52]. We term our proposed feature as *constant-Q transform based modulation spectral feature* (CQT-MSF). A 2-dimensional convolution neural network architecture (2-D CNN) is used to further refine the emotion information present in CQT-MSF feature. We compare the performance of CQT-MSF with mel-frequency spectral coefficients (MFSC) and show that the constant-Q non-linearity based auditory-cortical

features outperform the mel-scale non-linearity based features. We also investigate the performance differences obtained with auditory and cortical representations taken separately. We also highlight the striking similarity of CQT-MSF with the wavelet-based time-shift and deformation invariant coefficients, known as scattering transform coefficients [53]. Our main contributions in this work are as follows:

- This study proposes a new human auditory-cortical physiology based SER framework.
- We propose a modulation feature extraction technique using constant-Q filterbank over constant-Q spectrogram and analyse its relevance from vocal emotion perspective.
- We perform similarity analysis with another two-staged auditory-cortical feature representation: Scattering transform.
- We also perform explainability analysis to visually inspect different regions of CQT-MSF that weigh the most in prediction of a particular emotion class.
- The study further hints correlation between music training and emotion understanding by discussing the case of *Amusia* [54], and the possible analogy between modulation computed over CQT spectrogram and the cortex-level processing of sound in music trained individuals [55–60].

3. Early auditory processing: Constant-Q Transform (CQT)

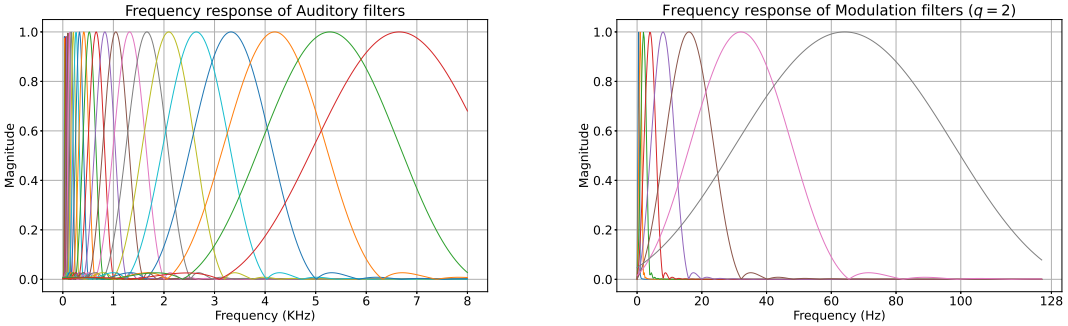
Our proposed features are based on the use of constant-Q filterbanks for both time-frequency (early auditory) and temporal modulation (cortical) based analysis of speech. In this section, we briefly discuss the CQT method of time-frequency representation. CQT uses constant *quality factor* (Q-factor) bandpass filters with logarithmically spaced center frequencies [61]. Mathematical formulation of constant-Q transform is given by,

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

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,

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

where f_k is the center frequency of a_k, f_s is the sampling frequency and $w(n)$ is the window function with length N_k. We use the standard *Hann* window in this work for CQT computation. The center frequencies of filters in constant-Q transform are



(a) Early auditory filterbank

(b) Modulation filterbank

Fig. 2: Auditory and Modulation filter banks used in CQT-MSF. The modulation filters shown here have scale factor (q) value 2.

spaced by the relation $f_k = f_{\min} 2^{\frac{k-1}{B}}$ where f_k is the frequency of kth filterbank, $f_{\min}$ being the frequency of the lowest bin and B the number of frequency bins used per octave of frequency. This binary logarithmic spacing leads to more frequency bins at lower frequencies, as compared to high frequencies, and hence provides higher frequency resolution at low frequencies [61]. In time domain, such filters can be given as truncated sinusoids (e.g., truncated with *Hann* window) with different lengths [62], given by,

$$N_k = \frac{q f_s}{f_k (2^{\frac{1}{B}} - 1)} \tag{3}$$

where q is the filter scaling factor. This scaling factor offers to change the time (and hence frequency) resolution of CQT bases without affecting B [62]. When compared with mel-based features, the mel-scale is also logarithmic in nature. However, mel-scale uses a decadic logarithm scale (or natural logarithm in some implementations), because of which, the emphasis on low-frequencies is not as prominent as CQT.

The computation of CQT, as described in Eq. 1, includes convolution of *atom* with every time sample of the input signal. However, the fast CQT computation algorithm [62] introduced a *hop length* parameter, which describes the number of samples the time window is shifted for next time frame CQT computation. The hop length is kept equal to integer multiples of $2^{\text{No. of octaves}}$ so that the corresponding signal frames at different frequencies do not fall out of alignment [62]. In CQT representation, the number of octaves is given by $\log_2 \frac{F_{\max}}{F_{\min}}$ [61] where $F_{\min}$ and $F_{\max}$ are the minimum and maximum frequency of operation, respectively. For CQT computation in this work, we use the *LibROSA*[1] toolkit [63] which follows all the computational details of the fast CQT implementation mentioned above.

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

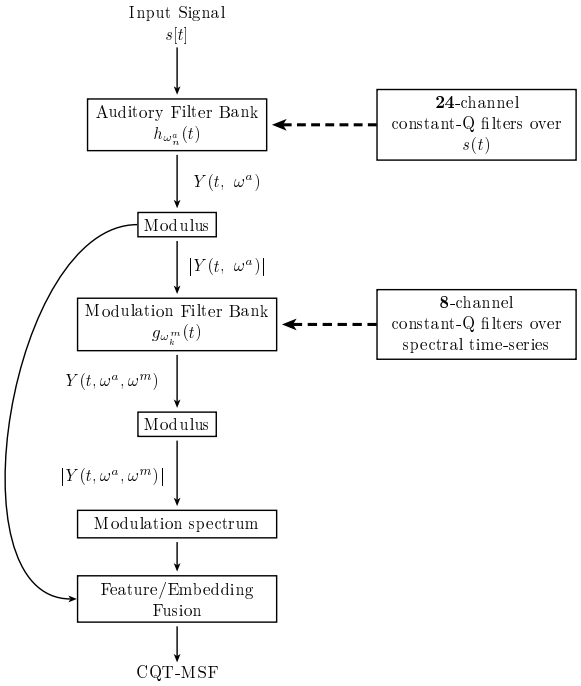


Fig. 3: Block diagram of proposed CQT-MSF feature extraction method. In figure, ω^a refers to acoustic frequency and ω^m refers to the modulation frequency.

4. Cortex-based processing: Modulation Spectrogram

Modulation spectrogram shows the temporal variation pattern of the spectral components in spectrogram. According to [51], speech signal is composed of two parts, the carrier, i.e., the vocal cords excitation, and the varying modulation envelope which is the result of changes in orientation of different vocal organs over time. The low-frequencies of the modulation envelope characterise slow variations of the complete spectral structure, which is known to encode most of the phonetic information [36, 64, 65]. Let $S(t,\ \omega)$ be the speech spectrogram. The temporal evolution of a frequency bin ω_o in $S(t,\ \omega)$, over time t, is a one-dimensional time-series. The spectral representation of this time-series $S(t,\ \omega_o)$ constitutes the modulation spectrum of frequency bin ω_o over T, where T is the spectrogram time window (with duration equal to window length N_k).

For speech, most of the modulation energy remains concentrated around 2-4 Hz range with peak at 4 Hz [51]. This makes 4 Hz to be considered as the syllabic rate of normal (*Neutral*) speech. Deviations from this rate generally result from infliction of noise or reverberation effects over speech [65, 66]. It is studied in SER literature that rate of speech is higher than *Neutral* class for high arousal emotions, such as, *Anger*, *Fear* and lower for low arousal emotions, such as *Sad*, *Boredom* [67]. Hence, this deviation of modulation energy peak from 4 Hz can be used for emotion discrimination over arousal scale.

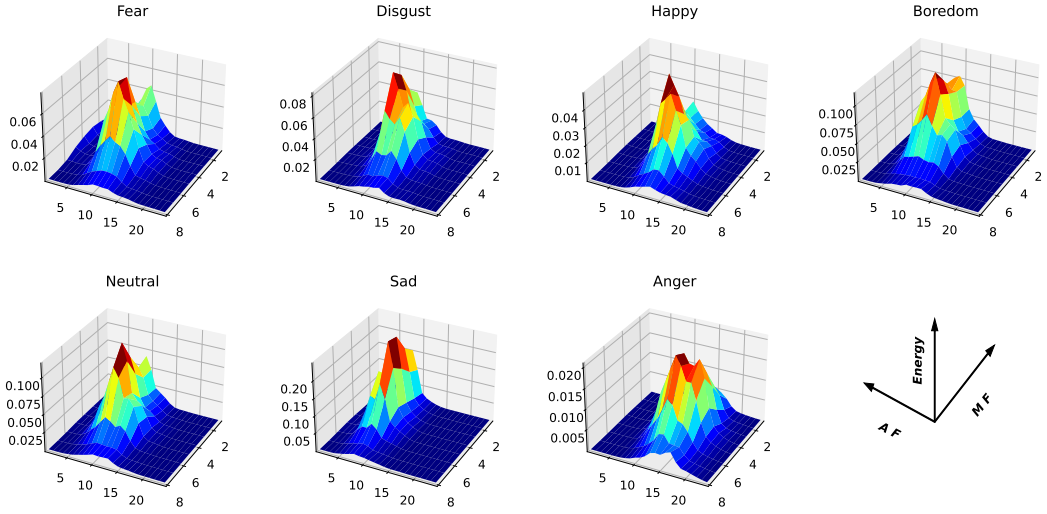


Fig. 4: Modulation spectral features (averaged over time) for different emotions in EmoDB. The '**M F**' refers to the modulation frequency channels and '**A F**' refers to the acoustic frequency channels. The modulation filters used in this analysis has filter scale (q) value 2.

4.1. Constant-Q based Modulation Spectral Features (CQT-MSF)

In this subsection, we compute modulations of CQT bins and combine them with CQT spectrogram to generate CQT-MSF. The first stage early auditory analysis (CQT spectrogram) in the CQT-MSF feature can be given as,

$$Y(t,\ \omega^a) = s(t) * h_{\omega^a_n}(t);\ \omega^a_0 \leq \omega^a_n < \omega^a_C, \tag{4}$$

where, $s(t)$ is the input speech signal, $h_{\omega^a_n}(t)$ is the impulse response of nth constant quality factor auditory filter with ω^a_n center frequency, C is the number of auditory filters and $Y(t,\ \omega^a)$ is the corresponding time-frequency representation. Fig. 2a shows the frequency response of different $h_{\omega^a_n}(t)$ used. For envelope extraction, modulus operation is applied over $Y(t,\ \omega^a)$, i.e., $|Y(t,\ \omega^a)|$. The resulting representation provides the temporal trajectories of different frequency bins in $Y(t,\ \omega^a)$.

For cortical analysis, the $|Y(t,\ \omega^a)|$ is passed through a modulation filterbank. The modulation spectrogram computed over time-frequency representation $|Y(t,\ \omega^a)|$, is given as,

$$Y(t, \omega^a, \omega^m) = |Y(t,\ \omega^a_n)| * g_{\omega^m_k}(t);\ \omega^m_0 \leq \omega^m_k < \omega^m_M,\ \omega^a_0 \leq \omega^a_n < \omega^a_C, \tag{5}$$

where, $g_{\omega^m_k}(t)$ is the impulse response of kth modulation filter with ω^m_k center frequency and M is the total number of modulation filters. Similar to the output of the first stage, we use the modulus of computed modulation spectrum coefficients computed over all

frequency bins of CQT spectrogram, i.e., $|Y(t, \omega^a, \omega^m)|$ [51]. Fig. 3 shows the block diagram of CQT-MSF feature extraction. The complete MSF includes concatenation of temporal modulations, computed using every modulation filter ($g_{\omega_0^m} \leq g_{\omega_k^m} < g_{\omega_M^m}$), of all frequency bins in the time-frequency representation, i.e., for $\omega_0^a \leq \omega_n^a < \omega_C^a$ bins where $C = 24$ auditory channels in our experiments. Regarding the properties of MSF, study performed in [31] report distinction between three different temporal modulation rates: slow, intermediate, and fast. The slow modulation rate is shown to roughly correspond to the syllable or speaking rate. Whereas the intermediate modulation rate appearing because of interharmonic interaction is shown to reflect the fundamental frequency of the signal. This shows the importance of temporal modulation for pitch representation, and hence, SER. Temporal modulation extracted by MSF represent tempo [68], pitch, and timber [52], all of which are related to emotion information in speech.

Fig. 4 shows the time-averaged CQT-MSF coefficients for utterances of different emotion classes of the EmoDB database. The MF and AF refer to the modulation and auditory frequency channels, respectively. In terms of modulation frequency, the highest peak in *Neutral* emotion is observed at 4 Hz modulation frequency with another peak around 0.5 Hz. Compared to *Neutral* class, low arousal emotions (*Boredom* and *Sad*) also have energies extending towards 0-4 Hz modulation frequency range. High arousal emotions (*Anger*, *Fear*) have peak around 4-8 Hz modulation frequency range. In contrast, *Happy* also has a peak at 4 Hz similar to *Neutral*. Similarly, *Disgust* also shows a peak at 4 Hz followed by another peak at 2 Hz. From AF perspective, *Anger* emotion shows peak at high frequencies (high AF), whereas, in *Sad* low auditory frequencies are more dominant. For remaining emotions, auditory energy distribution extends almost similarly over mid-auditory frequencies. This analysis shows the higher emotion discrimination potential of combined MF and AF channels as compared to only AF channel-based representation.

To further analyze the discriminative potential of modulation spectrum features, we perform F-ratio analysis between the time-averaged modulation features of various emotion classes and the *Neutral* class of the EmoDB database [69, 70]. Fig. 5 shows the 3-D projection of F-ratio values in AF-MF plane. Different auditory and modulation bins show varying discriminative characteristics for different emotions. For every emotion class, F-ratio peaks are observed at low MF bins showing their potential for emotion discrimination. Similarly, low AF also shows high F-ratio for every class except *Sad*. High arousal emotions (*Anger*, *Happy*, *Fear, etc.*) in general show greater F-ratios at high MF bins as compared to low arousal emotions (*Sad*, *Boredom*). The highest F-ratio value with respect to *Neutral* is observed for *Anger* and lowest for *Boredom* emotion class. For *Anger* and *Happy* classes, low AF bins exhibit higher discriminative characteristic. *Disgust* shows F-ratio peaks over wide range of AF and corresponding MF bins. In *Fear*, F-ratio peaks are observed at low and high AF values with gradual slope towards increasing MF bins. The presence of moderately higher F-ratio values at MF bins is a result of increase in speaking rate in high arousal emotions. *Boredom* class has lowest F-ratio values, mostly focused at low AF and low MF bins, whereas, *Sad* shows higher discrimination w.r.t. *Neutral* over low and high AF and MF bins. Lower F-ratio values for *Boredom* also indicate its similarity in characteristics with *Neutral* class. The F-ratio analysis again shows the higher discrimination potential of joint AF and MF bins, with respect to *Neutral* emotion.

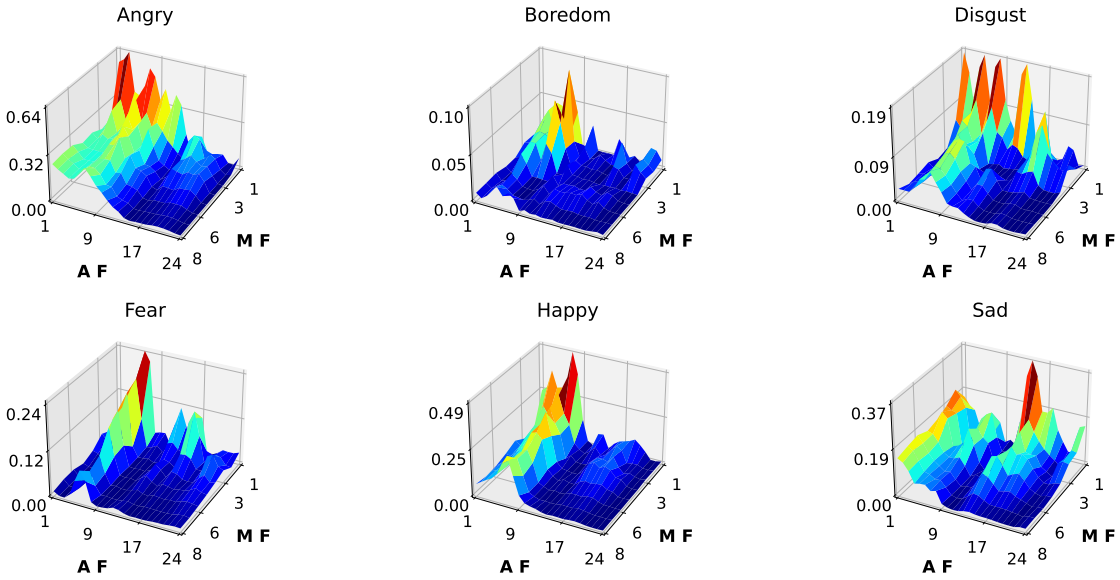


Fig. 5: F-ratio values of different auditory frequency (AF) and modulation frequency (MF) bins of modulation spectrum features computed over EmoDB database. The F-ratio is calculated over time-averaged modulation spectrum features (MSF) between *Neutral* and every other emotion class. The modulation filters used in this analysis have filter scale (q) value 2.

4.2. Comparison between CQT-MSF and Scattering Transform

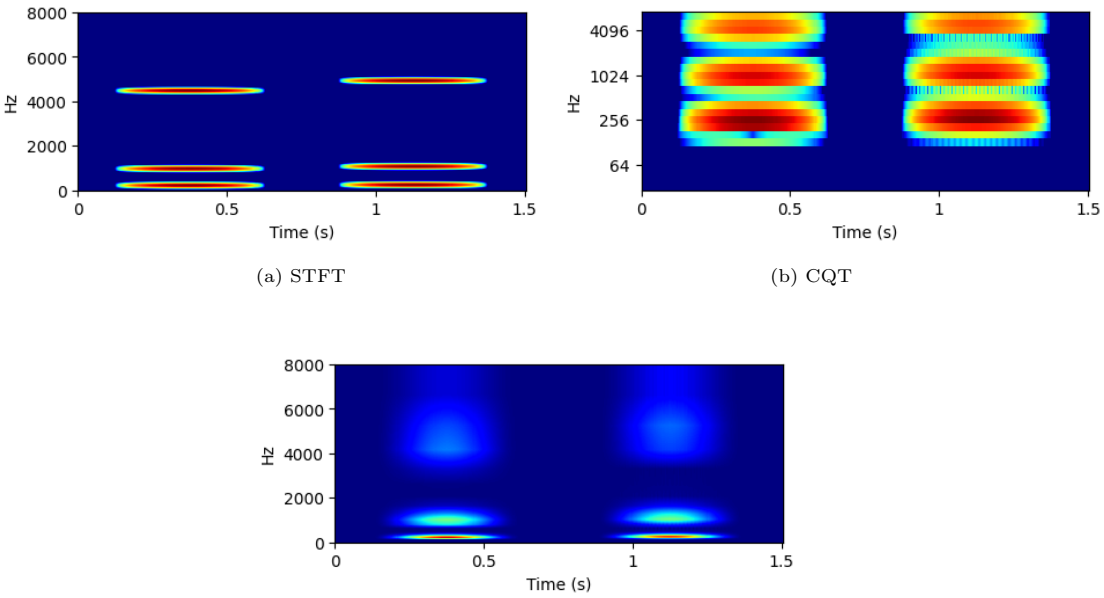
Our proposed CQT-MSF feature, combined with standalone CQT, has striking similarity with *scattering transform* feature representation of 1-D signals. Authors in [53] compute scattering coefficients of 1-D signals and show their characteristic invariance against temporal shifts and deformations. The features (or coefficients) are computed by convolving the signal with a set of predefined filter kernels. The feature extraction process includes the following steps: 1) Scalogram computation by passing the signal through a bank of wavelet filters. 2) Passing the obtained time-series of frequency bins in scalogram through another set of wavelet filterbank to obtain modulation spectrogram. 3) Introduce stability to deformations by low-pass filtering the signal, scalogram and modulation spectrogram coefficients. The scattering transform coefficients are mathematically described as,

$$S_{J_2}x(t) = U_2x * \phi_{2^J}(t) = \int U_2x(u)\phi_{2^J}(t-u)du \tag{6}$$

where,

$$U_2x = U[\lambda_2]U[\lambda_1]x = ||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|. \tag{7}$$

Here, x is the 1-D signal, $\phi_{2^J}(t)$ defines the averaging low-pass filter with scale 2^J, ψ_{λ_N} describes the Nth layer complex *Morlet* wavelet filterbank (layer 1 are scalogram



(a) STFT

(b) CQT

(c) CQT with linear frequency scale

Fig. 6: Visual description of deformation stability of STFT and CQT. a), Signal $x(t)$ (left) and its deformed version $x`(t)$ (right). b), CQT representation of the same signal and its deformed version. c) CQT of the original and deformed signal projected on linear frequency scale. Figure taken with permission from [46].

coefficients and layer 2 constitutes modulation coefficients) and operator '$*$' is the convolution operator. Scattering coefficients are found useful in various speech and audio processing domains, e.g., speech recognition [53], speaker identification [71], urban and environmental sound classification [72], etc. In [73], scattering coefficients also showed improvement in SER performance over mel-frequency cepstral coefficients (MFCCs).

In our proposed CQT-MSF feature, the CQT time-frequency representation is similar to the first layer scalogram coefficients computed by scattering transform. Similarly, the MSF computed over CQT is similar to the modulation spectrogram computed over scalogram in second layer of scattering transform. Also, CQT follows the same constant-Q non-linearity as followed by the filterbanks in both first and second layers of the scattering transform. However, the averaging performed in scattering coefficients to obtain invariance to time-shift and deformations is absent in CQT-MSF. The design parameter (e.g., bandwidth) of the low-pass filter which performs this averaging in scattering transform is manually selected depending upon the input signal characteristics. To address the absence of time-shift invariance, we employ a 2-D convolutional neural network over the computed CQT-MSF. The employed CNN architecture includes multiple layers with different filter scale values in different layers. According to [74], convolutional neural networks inherently exhibit invariance in vertical direction (direction of network depth) which mainly appears due to feature pooling. Hence, a generic CNN architecture with pooling layers can learn to apply the required averaging and obtain the required time-

shift invariance characteristic.

Regarding sensitivity to deformation, authors in [74] and [75] prove that convolutional feature extractors provide inherent but limited deformation stability. The extent of this stability depends upon the deformation sensitivity of the input signal. Signals which are slowly varying or band-limited are more deformation insensitive than signal with sudden changes or discontinuities [75]. As the CQT also provides a non-uniform filterbank representation as provided by mel-filters, similar stability to temporal deformation can be assumed in CQT filterbank as well. Fig. 6 shows the deformation stability of STFT, CQT and CQT with linear frequency scale for a signal $x(t)$ deformed by a factor ϵt (i.e, $x'(t) = x(t - \epsilon t) = x((1 - \epsilon)t)$) [53]. The upward shift of spectral response in STFT appears due to the 'ϵt' term, leading to instability to deformation imposed by ϵt. However, in CQT, spectral responses of deformed signal do not show any major frequency shift. Instead, the deformed signal well-overlaps with the original signal because of higher filter bandwidth at higher frequencies. This shows that CQT is indeed deformation stable as compared to STFT. The linear frequency CQT plot is given for comparison of STFT and CQT over linear frequency scale, hence confirming the deformation stability of CQT in both linear and non-linear frequency scales.

Therefore, convolutional neural network layers can be used to inherently provide the required time-shift and deformation invariance for better emotion-rich representation at its output. We hence write the feature extracted by convolution layers of our employed DNN model as,

$$S = F(||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|) \tag{8}$$

where, $F(.)$ is the function estimated by 2-D convolution layers, and $\psi_{\lambda 1}$ and $\psi_{\lambda 2}$ corresponds to the filterbanks $h_{\omega_n^a}(t)$ and $g_{\omega_k^m}(t)$ used in the CQT-MSF generation (Fig. 2). Another point of dissimilarity is the difference between the basis functions used in the filterbank of scattering transform and CQT-MSF. The former uses *Morlet* wavelets, whereas, the latter employs sinusoids multiplied with *Hann* window function. The ripples observed in the frequency response of the filters in Fig. 2 is because of the small spectral leakage in the *Hann* window.

5. Experimental Setup

5.1. Database Description

For analysis of CQT-MSF and its comparison with mel-scale features, we perform experiments with two different speech corpora. We use Berlin EmoDB and RAVDESS datasets which are most widely used and publicly available.

5.1.1. Berlin Emotion Database (EmoDB)

Berlin Emotion Database [76] contains acted emotional speech recordings of 10 professional artists (5 female and 5 male). The actors speak ten emotionally neutral and phonetically rich sentences in German language. Seven different emotion categories are used in the database: *Anger, Happy, Fear, Sad, Boredom, Disgust,* and *Neutral*. To

evaluate the authenticity of recordings, listening test was performed by 20 subjects. A total of 800 utterances were recorded but only 535, having more than 80% recognition rate and 60% naturalness, were finally selected. Our choice of this database is explained by its diligent recording setup, popularity in SER domain [8, 12, 13, 37, 77–80] and its free availability.

5.1.2. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The RAVDESS database [81] contains acted utterances of 12 male and 12 female artists speaking English language. A total of 7536 clips were recorded in three different modalities, namely audio-only, video-only, and audio-video, out of which the audio-only modality contains 1440 spoken utterances from all speakers. The database includes eight different emotion categories (*Happy, Anger, Sad, Neutral, Disgust, Calm, Surprised,* and *Fear*) with two intensity levels, strong and normal. Recorded clips were evaluated by 319 subjects out of which 247 tested the validity and 72 evaluated test-retest reliability of recordings. An average of 60% accuracy was obtained in validity test over recordings of all emotions. Up-to-date design and inclusion of an extensive emotion set with varying intensities make this an important database for SER.

To further increase the diversity in training data, five-fold data augmentation following the x-vector *Kaldi*[2] recipe is used [82]. The augmented data involves adding additive and reverberation noises over clean speech samples. The RAVDESS database is downsampled to 16 kHz before data augmentation and feature extraction.

5.2. Parameter settings for feature extraction

We compare the performance of proposed CQT-MSF features with baseline CQT and MFSC. The different parameter values used for CQT and MFSC are based on our preliminary comparison of the two methods [47]. For CQT computation, we select the minimum frequency value F_{min} to be 32.7 Hz and F_{max} equal to the Nyquist frequency. This provides a total of eight octaves over complete frequency range. Every frequency octave contains three bins which provides a total of 24 frequency bins over complete frequency range (F_{min} to F_{max}). The obtained CQT representation corresponds to 24-channel early auditory stage feature extraction. Another important parameter in CQT computation is hop length which is the number of samples the observation window advances between successive frame-shifts. We keep the hop length value fixed at 64. For cortical analysis, the above-mentioned CQT representation is passed through another set of constant-Q filterbank, referred to as modulation filterbank, as described in Section 4. We use an 8-channel modulation filterbank with center frequencies ranging from 0.5 to 64 Hz covering a total of eight frequency octaves.

For a fair comparison with CQT and CQT-MSF, similar modification in parameter values of MFSC is applied. We use 24-filter bank for MFSC computation over STFT with 512 frequency bins. The frame size is fixed to 320 samples with hop length of 64 samples over 16 kHz sampling frequency. We also compute the modulation spectrum coefficients by using MFSC time-frequency representation for comparison with CQT-MSF features.

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

We refer to these as MFSC-MSF features. We use the same *LibROSA* toolkit for MFSC feature generation.

5.3. Evaluation Methodology

Unlike other speech processing tasks, such as automatic speech recognition (ASR), automatic speaker verification (ASV), the SER lacks standardization of evaluation methodology for performance benchmarking on publicly available datasets. This leads to a wide variation in results across different works reported in the literature. Some examples of differences in evaluation protocol are the use of some selected emotions from the databases, the choice of performance evaluation metric, the selection of cross-validation strategy, the differences in the selected emotion classes, etc. Because of these reasons, meaningful comparison of obtained results with those reported in the literature becomes inaccurate, if not impossible, in SER research.

We adopt a leave-one-speaker-out (LOSO) cross-validation strategy for evaluation and benchmarking. The databases are divided into train/validation/test groups with every group containing disjoint speakers. The test and validation group contain utterances from one speaker and the remaining speakers are kept for training. This keeps the total number of train/validation/test sets equal to the number of speakers in every database. The final performance over a database is reported by averaging the performance metrics obtained for every train/validation/test group. For SER, speaker-dependent testing is known to fair better than speaker-independent testing [83]. However, speaker-independent sets of the database eliminate the chances of the trained classifier being biased towards a set of speakers and also simulates the real-world scenario in a better way. Although LOSO cross-validation is computationally expensive, due to small database sizes in SER, the increase in complexity can be safely ignored. Also, keeping a single speaker for testing allows more training data to be available which is essential with small databases.

5.4. Classifier Description

In this work, we use two different machine learning frameworks to evaluate performances of studied features: (1) convolutional neural network with fully connected layer for emotion classification (termed henceforth as DNN). (2) Convolutional layers to extract emotion embeddings and SVM to classify embeddings into emotion classes (termed as DNN-SVM). Our selection of these is inspired from the success of embeddings based networks [12, 82, 84] and fully DNN-based frameworks [40] in speech processing. Performance evaluation over these also enables us to compare the SER efficiency of the two DNN frameworks.

The DNN-SVM framework comprises of two parts: embedding extraction from convolutional layers of the trained model used in the DNN framework, and final classification using SVM. The SVM is trained and tested over the embeddings extracted from the trained DNN model. We extract embeddings at the output of global average pooling (GAP) layer, placed after the final convolutional layer. These embeddings are processed with an SVM back-end for performance evaluation. In SVM model, we empirically select the value of regularization parameter C and the expanse/width of the *radial basis*

Table 1: The parameters of CNN architecture for SER. The number of 2D-Conv layers and the kernel sizes are inspired from the x-vector TDNN architecture [82]. Maxpooling applied after every 2D-Conv layer provides time and frequency invariant feature representations.

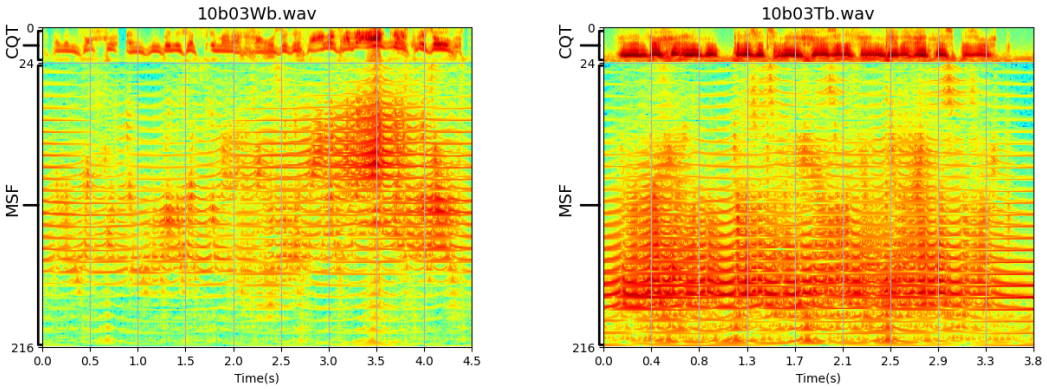
Layer	**No. of Filters**	**Height (Frequency)**	**Length (Time)**
2-D Conv	128	5	5
Maxpool	-	2	1
2-D Conv	128	3	3
Maxpool	-	2	1
2-D Conv	128	3	3
Maxpool	-	2	1
2-D Conv	128	1	1
Maxpool	-	2	1
Global Average Pool (GAP)	-	-	-
Fully Connected	64	-	-
Softmax	#Classes	-	-

function kernel (parameter γ) to 1 and 0.001 respectively [85].

In the DNN framework, to train and validate the model, the speech utterances from every database are chunked into segments of 100 frame length with 50% overlap across consecutive frames. With 64 samples hop and 16 kHz sampling rate, this corresponds to 400 ms speech duration. Our choice of 400 ms is based on the reports in SER literature which explain that segment length greater than 250 ms contains required information for emotion prediction [12]. However, for testing, complete utterances are used to test model performance. This is done as the labels provided to emotion speech recordings are over complete utterances and not over segments. This approach also leads to increase in available data samples for training. Similarly, in DNN-SVM framework, the train embeddings are generated over segments of speech utterances with 100 frame length (400 ms), whereas, test embeddings are generated from complete utterances.

Table 1 describes the DNN architecture employed in this work. We use cross entropy optimizer with learning rate value of 0.001 with 64 batch size and dropout value of 0.3 applied over only the fully connected (FC) layer. The model is trained for 50 epochs and the version with the best performance on validation set is used for testing.

As the convolution layers accept input in 2-D (time-frequency) form, we use two different strategies to combine CQT/MFSC time-frequency representation with their MSF features. In the first method, we directly concatenate the CQT/MFSC with its corresponding MSF features over the frequency axis. This leads to a representation with time frames in x-axis and early auditory frequency bins, followed by modulation frequency bins corresponding to every auditory bin, placed in succession over y-axis. Fig. 7 shows



(a) CQT-MSF of *Anger* emotion

(b) CQT-MSF of *Sad* emotion

Fig. 7: Logarithm of feature fusion of CQT and modulation spectral features, i.e., CQT-MSF, extracted from CQT over utterances taken from EmoDB database. The first 24 bins on y-axis correspond to the CQT spectrogram. Bins that follow include stacking of 8 modulations bins corresponding to every auditory bin. The total number of bins then becomes, 24 auditory bins + 24 auditory bins × 8 modulation bins for every auditory bin = 24 auditory bins + 192 MSF bins = 216 total bins on y-axis).

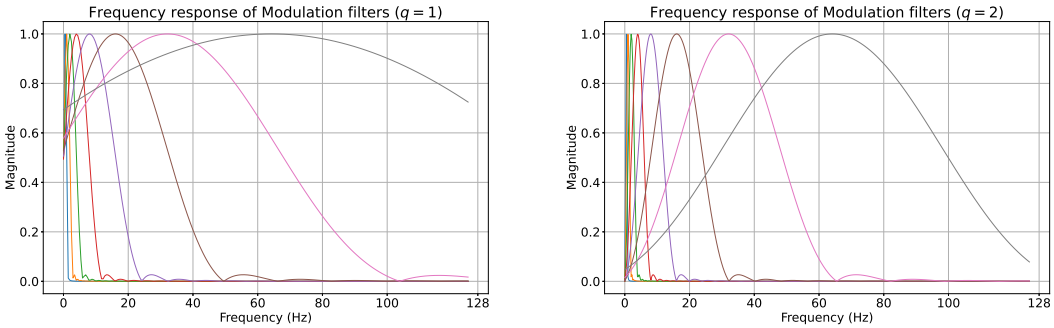
the 2-D representation obtained with concatenation of CQT/MFSC with the corresponding MSF. In second approach, to better combine the information from time-frequency and modulation features, we use an embedding fusion based DNN architecture. The architecture consists of two parallel but similar branches of convolutional and GAP layer, followed by a common FC and softmax layer. For both feature fusion and embedding fusion, the embeddings are extracted from the GAP layer of DNN model.

5.5. Evaluation Metrics

For performance evaluation, we use accuracy and UAR metrics. We chose these metrics owing to their popularity in SER and also for better comparison of results with the literature. Accuracy is defined as the ratio between the number of correctly classified utterances to the total number of utterances in test set. According to [86], the UAR metric is given as:

$$\text{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}} \tag{9}$$

here, A is called the contingency matrix, A_{ij} refers to the number of samples in class i classified as class j, and K is the total number of classes. As accuracy is considered *unintuitive* for databases with uneven samples across different classes, we use UAR to measure the validation set performance of the DNN model and to select the best performing model over the set of epochs.



(a) Filter scale $q = 1$

(b) Filter scale $q = 2$

Fig. 8: Modulation filter banks for different values of filter scale factor q. Filters with $q = 1$ have same center frequencies but higher bandwidth than filters with $q = 2$.

Table 2: Comparison between different filter scale (q) values of modulation filters for experiments performed with feature-fused CQT-MSF over EmoDB database. Given values are in percentages.

Classification	$q = 1$		$q = 2$		$q = 3$	
Framework	**Accuracy**	**UAR**	**Accuracy**	**UAR**	**Accuracy**	**UAR**
DNN	76.97	68.25	70.79	64.93	69.77	64.27
DNN-SVM	79.86	76.17	79.50	77.00	74.28	70.91

6. Results & Discussion

6.1. Performance Comparison of Different Features

First, we experimentally optimize the constant-Q filters used for CQT-MSF computation by varying the value of filter scaling factor (q). Fig. 8 shows the frequency response of modulation filter banks with $q = 1$ and 2. Since q affects the time resolution of filters as given in Eq. 3, filters with $q = 1$ are wider (have higher bandwidth) which leads to clipping of the filter frequency response at low frequencies. This leads to inclusion of zero-frequency (or DC) components as well. Also, because of greater overlap between filters, the generated filter outputs have higher redundancy. Increased redundancy helps convolutional layers to better extract required emotion correlation among modulation frequency bins. With $q = 2$, the filter responses remain limited inside the frequency range providing a less redundant filterbank structure as shown in Fig. 8b. Table 2 reports the difference in results obtained for modulation features computed with different values of scaling factor q. Filterbank structure with $q = 1$ outperforms the arrangement with $q = 2$ and 3. Hence, we select modulation filters with $q = 1$ for further experiments. We perform optimization and detailed experimentation of q over only EmoDB database due to its small size and similar trends with other databases.

Table 3 shows the performance of time-frequency representations combined with their corresponding modulation features for different classification frameworks over EmoDB

Table 3: Performance comparison of combined early auditory and cortical features over EmoDB database. Filter scale (q) value of 1 is used in modulation filterbank for MSF computation. Given values are in percentages.

Features	**DNN**		**DNN-SVM**	
	Accuracy	**UAR**	**Accuracy**	**UAR**
CQT-MSF (Feature Fusion)	76.97	68.25	79.86	76.17
CQT-MSF (Embedding Fusion)	77.99	71.74	79.32	75.48
MFSC-MSF (Feature Fusion)	61.45	55.23	69.02	64.45
MFSC-MSF (Embedding Fusion)	60.80	57.92	67.74	64.89

Table 4: Performance comparison of early auditory and cortical features taken separately over EmoDB database. Filter scale (q) value of 1 is used in modulation filterbank for MSF computation. Given values are in percentages.

Features	**DNN**		**DNN-SVM**	
	Accuracy	**UAR**	**Accuracy**	**UAR**
MSF (Computed over CQT)	72.67	66.32	78.73	76.33
MSF (Computed over MFSC)	59.35	53.44	65.24	60.54
CQT	71.77	64.91	76.74	73.21
MFSC	61.62	57.32	66.18	62.58

Table 5: Feature performance comparison over different databases with DNN-SVM framework. Given values are in percentages

Database	**MFSC**		**CQT**		**CQT-MSF (Feature Fusion)**		**Chance level performance**
	Accuracy	**UAR**	**Accuracy**	**UAR**	**Accuracy**	**UAR**	**Accuracy**
EmoDB	66.01	61.75	76.74	73.21	79.86	76.17	14.28
RAVDESS	36.94	36.16	48.68	44.64	52.24	48.83	12.5

database. The DNN-SVM classification framework outperforms DNN framework for every feature and over both performance metrics. This observation is counter-intuitive

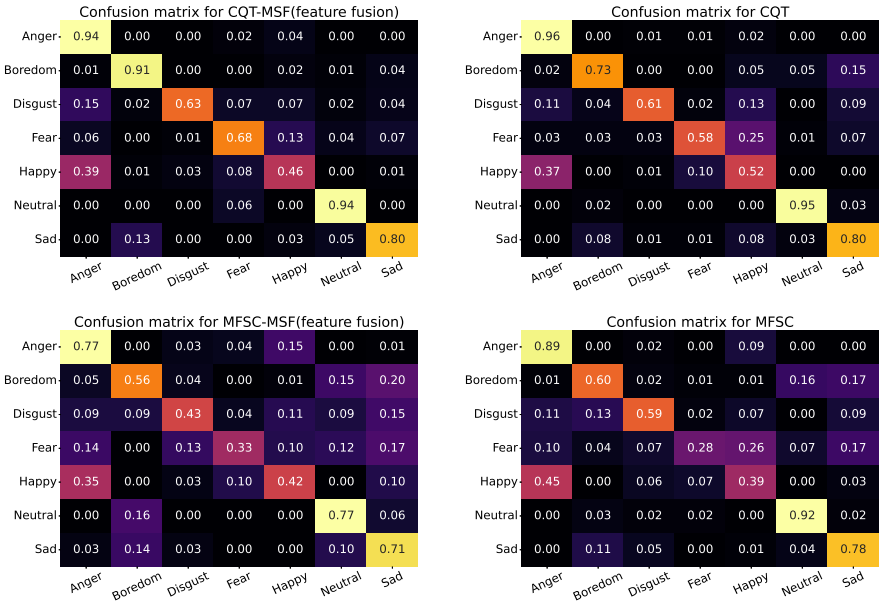


Fig. 9: Confusion matrices for CQT-MSF (feature fusion), MFSC-MSF (feature fusion), CQT, and MFSC features. CQT-MSF shows best comparative performance in classifying different emotion classes of EmoDB database. Matrices are computed over DNN-SVM framework owing to its greater performance.

as the activations extracted from the same convolutional layer in DNN-SVM and DNN framework, end up performing better with SVM at back-end but not with fully-connected layer at the back-end. Regarding the two different feature fusion types, there is no specific pattern, in terms of performance improvement, among the classification frameworks. With CQT-MSF, feature fusion outperforms embedding fusion over DNN framework, whereas, embedding fusion outperforms feature fusion for DNN-SVM framework. For MFSC-MSF, the trend is opposite to that of CQT-MSF fusion results. However, CQT-MSF in both fusion styles performs better than MFSC-MSF.

To further analyse the contribution of early auditory and cortical features taken separately over SER, we perform experiments to analyse the performance of standalone MSF extracted over CQT/MFSC (without any type of fusion). From the results in Table 4, cortical features (standalone MSF over CQT) outperforms standalone CQT. However, the same is not true for mel-scale based features. The MSF computed over MFSC shows poor performance as compared to standalone MFSC. This questions the usability of temporal trajectories of the mel-scale time-frequency representation for emotion classification. The inferiority of MFSC against CQT is also indicated by direct comparison between CQT and MFSC. CQT outperforms MFSC in both DNN and DNN-SVM classification frameworks. Among different CQT feature types, MSF computed over CQT assumes very similar performance in contrast to both fusion types of CQT-MSF. This phenomenon describes the higher emotion relevance of the temporal modulations of the low-frequency regions emphasised by CQT.

Fig. 9 shows the confusion matrices for different features over EmoDB database. We choose only the feature-fused CQT-MSF and MFSC-MSF for comparison, owing to their improved performance as compared to embedding fusion in DNN-SVM framework. Even though CQT-MSF is comparatively better at classifying different emotion classes, some instances of *Happy* and *Disgust* are confused with *Anger*, and that of *Fear* are confused with *Happy*. The highest misclassification is in *Happy-Anger* emotion pair which have similar arousal but opposite valence characteristic. This observation is found to be consistent among various SER works [12, 40, 87] and can also be related to the similar F-ratio characteristics of *Anger* and *Happy* in Fig. 5. Among low-arousal classes, some confusion in *Sad-Boredom* is also visible in CQT-MSF. As prosody features are less effective in valence discrimination [88], the confusion among pairs with similar arousal but opposite valence characteristics can be attributed to higher emphasis over pitch in constant-Q scale based spectral representation. However, modulations computed across constant-Q scale reduce this confusion, which is evident from the comparison between standalone CQT and CQT-MSF. In standalone CQT, confusion among both low and high arousal emotions is higher and appears among multiple emotion classes (e.g., *Disgust* and *Fear* with *Happy*). In MFSC and MFSC-MSF, as compared to CQT-based features, the misclassification is more prominent across multiple classes. Unable to emphasise speech prosody, the emotion classification ability of MFSC over arousal scale is inferior to CQT-based features (e.g., increased confusion of *Boredom* with *Neutral*, and *Fear* with *Sad*). Also, modulations of MFSC are computed with more focus on high and mid speech frequencies and less focus on prosody at low frequencies, further deteriorating the performance.

At the end, we compare the SER performance of CQT-MSF feature with the scattering transform coefficients. As scattering network is also a deep convolutional network, its comparison with our proposed CQT-MSF with DNN-SVM classification framework shows the superiority of automatic feature extraction and required time and frequency-shift invariance learning for SER. The scattering coefficients are computed using the similar train/test strategy. The training speech utterances are chunked to 400 ms segments with 50% overlapping, whereas testing utterances are used as is. Following our experiments in [73], the Q-factor value (Q) for first layer coefficients is chosen as Q = 5. As the training segment size is fixed to 400 ms (6400 samples at 16 kHz), the maximal wavelet length or averaging scale (T) is kept 4096 samples for both training and validation. However, since we use complete utterances in test, the duration N for testing is empirically fixed to 51000 samples (3.18 seconds at 16 kHz). Longer utterances are chopped to contain only 51000 samples, whereas shorter frames are zero-padded. We obtain **72.67**% accuracy and **69.8**% UAR with scattering coefficients outperforming MFSC, and indicating the requirement of time and deformation stability in SER. However, the performance is inferior to that with the proposed CQT-MSF (especially with feature-fusion). The superiority of CQT-MSF shows that deep networks learn to provide better time and deformation stability, as described in Section 4.2, while extracting the emotion relevant information from multiple convolution layers. Although scattering transform also involves convolutions and averaging for stability, it is performed using fixed kernels which are not automatically learned/optimized to improve performance.

Table 5 shows the results obtained with different features over EmoDB and RAVDESS databases. The proposed CQT-MSF feature outperforms other features over RAVDESS database as well. This explains the suitability of CQT-MSF features or two stage auditory analysis for SER over different databases. Compared to EmoDB, the relative performance improvement with CQT-MSF, CQT over MFSC is higher in RAVDESS database.

6.2. Comparison with Related Works

In this subsection, we compare our obtained results with related works in SER. Among SER literature, different strategies are used to evaluate system performances, for example, use of different databases, number of emotion classes used in the databases, different evaluation methodologies, etc. These differences make direct comparison of SER works difficult. The evaluation methodology, in terms of, cross-validation scheme, train/test split, speaker dependent/independent testing, etc. are found to differ substantially in SER literature. Lack of reproducible research in SER domain also leads to uncertainty, leading to difficulty in comparison. Hence, a comparison made with other relevant SER literature can not be considered accurate.

Due to the above mentioned issues, to justify our obtained results, we implement different studies from the literature by using our proposed CQT-MSF feature and experimental framework. Table 6 shows the list of selected works and the corresponding performances obtained. Section 5.3 and 5.4 of the manuscript describe the experimental framework employed in the studies listed in the table. Our choice of selected works is based on the use of modulation spectrogram related features, and use of advanced neural network architectures (e.g., contextual long short-term memory (LSTM), multi-head attention, ResNet architecture, multi-time-scale kernel, etc.). Below we briefly describe the details of selected works.

- Study performed by Avila et al. (2021) proposes feature pooling techniques over different measures of modulation spectral features for dimensional emotion recognition. For performance comparison, we compute the same modulation spectral measures over CQT representation, unlike Avila et al. which uses GMT representation, and show the results on our experimental framework. Feature-pooling schemes reported in the study are skipped to maintain similarity in comparison as our framework does not include handcrafted pooling operations.

- Aftab et al. (2022) use mel-frequency cepstral coefficients (MFCCs) over a fully convolutional neural network architecture with parallel paths of different kernels sizes followed by stacked convolutional layers for SER with impressive reported SER performance. We use the GitHub implementation[3] of the state-of-the-art architecture with our experimental framework and CQT-MSF feature for performance comparison.

[3] `https://github.com/AryaAftab/LIGHT-SERNET`

Table 6: Performance comparison with selected works. **Boldface** values show the best obtained results.

References	Brief Description (Feature & Classifier)	Performance (in %) EmoDB Acc.	EmoDB UAR	RAVDESS Acc.	RAVDESS UAR
Avila et al. (2021)	Different modulation spectral measures. DNN classifier.	55.40	46.83	37.77	29.10
Aftab et al. (2022)	MFCC. Parallel path fully convolutional network.	73.37	67.21	48.68	44.51
Liu et al. (2022)	Interspeech 2009 feature set. bc-LSTM + Multi-head attention.	53.45	47.49	24.58	21.79
Gerczuk et al. (2021)	Mel-spectrogram. Convolutional ResNet.	76.34	71.07	52.56	49.46
Guizzo et al. (2020)	Spectrogram. Multi-time-scale kernel based CNN.	63.01	58.03	40.27	36.46
Parra-Gallego and Orozco-Arroyave (2022)	x-vectors, i-vectors, INTERSPEECH 2010 feature set, articulation, phonation, and prosody features. SVM classifier.	50.21	42.74	46.59	43.68
This work	GMT-MSF feature. DNN-SVM framework.	77.05	74.35	**54.05**	**51.16**
This work (Proposed method)	CQT-MSF feature. DNN-SVM framework.	**79.86**	**76.17**	52.24	48.83

- We also select the state-of-the-art study performed by Liu et al. (2022) for multi-modal emotion recognition in our work. As our primary focus is emotion recognition from speech, we use only the bidirectional-contextualised LSTM (bc-LSTM) with multi-head attention block used for speech modality in [90] with our experimental framework and databases. As emotion information spreads temporally across utterances, temporal pattern extraction architectures like LSTM, attention, are selected to compare with handcrafted temporal modulation feature, i.e., CQT-MSF.
- Study reported by Gerczuk et al. (2021) employs an adapter ResNet architecture for multi-corpora SER scenario. As our study does not include mixing different corporas, we select the reported ResNet architecture without the adapter module but with CQT-MSF feature for performance comparison. We use the open-source GitHub implementation[4] of the model.
- Study performed by Guizzo et al. (2020) uses a multi-time-scale convolutional kernel based front-end which employs multiple temporally resampled versions of the

[4] `https://github.com/EIHW/EmoNet`

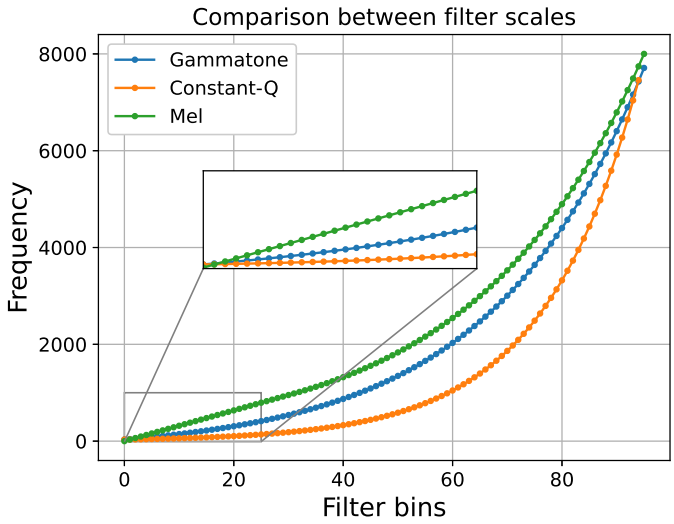


Fig. 10: Comparison between different non-linear time-frequency representation scales with filterbank center frequencies (shown by dots). The zoomed-in portion shows the difference between the non-linearities for emotion-salient low frequency filter bins. For better visibility of differences among the scales, we plot every scale with 96 frequency bins.

original convolution kernel and parallelly perform convolution with every version. We utilize this method by employing multi-time-scale convolution layer as front-end over our DNN framework (mentioned in Section 5.4). We use the available open-source GitHub implementation[5] of multi-time-scale convolution layer. Similar to CQT-MSF, the multi-time-scale kernels also focus on the extraction of speech temporal patterns for emotion recognition.

- We select the feature set based study by Parra-Gallego and Orozco-Arroyave (2022) for notably high performance reported on the RAVDESS database. The work includes combination of x-vector, i-vector, Interspeech 2010 paralinguistic (IS10) feature set, and articulation, phonation and prosody features extracted from *Disvoice*[6] framework for emotion classification with SVM as classifier. As the study report that the combination of only x-vector, IS10, and Disvoice framework provide the best performance, we use the same shortened feature set with SVM classifier, but on LOSO cross-validation framework. To maintain similarity, we use complete utterances (no segmentation) for both training and testing in this particular implementation. Note that study does not integrate our CQT-MSF and is rather based on the original implementation in the paper.

- As several modulation feature based SER works use gammatone-scale to generate a time-frequency representation over which modulations are computed [37, 38, 41],

[5] `https://github.com/ericguizzo/multi_time_scale`
[6] `https://github.com/jcvasquezc/DisVoice`

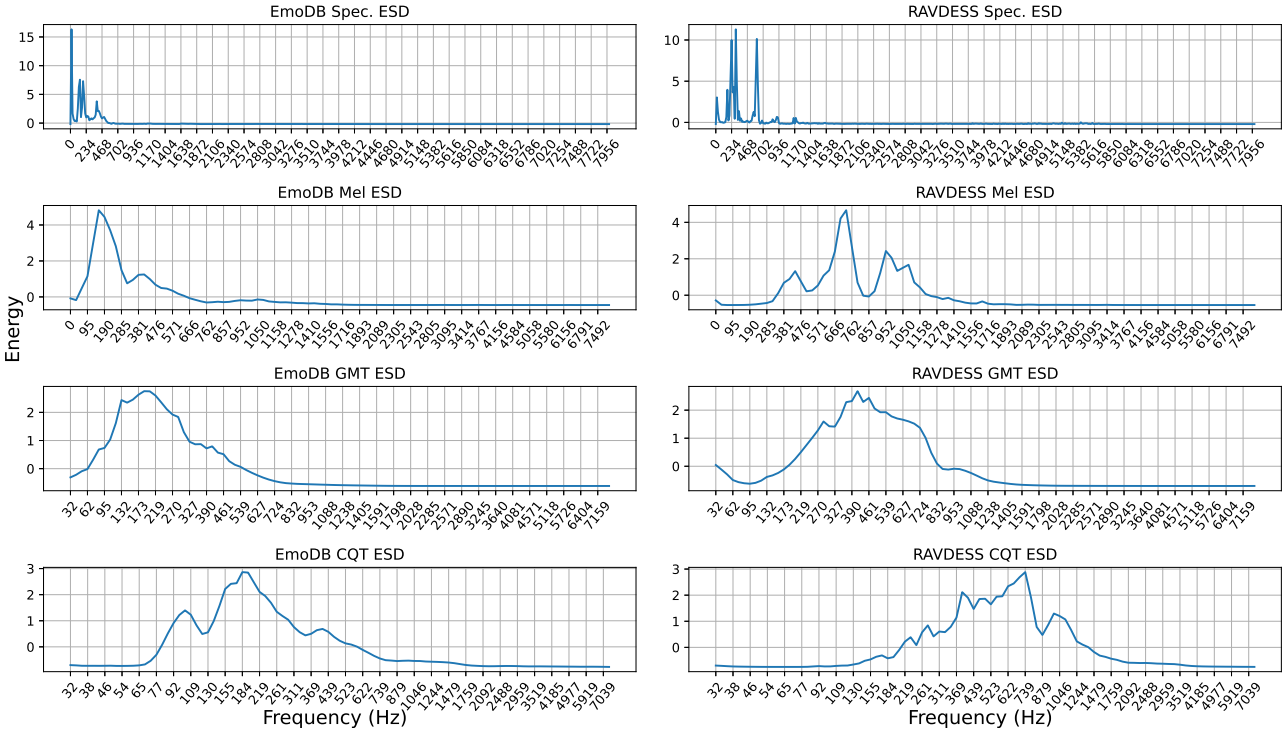


Fig. 11: Average energy spectral density (ESD) computed over all utterances of EmoDB and RAVDESS. The spectrogram (Spec.), MFSC, CQT, and GMT ESD corresponds to the energy of corresponding coefficients average across all utterances. For better visibility of differences among the features, we plot every ESD with 96 frequency bins.

we compare the performance of CQT-MSF with gammatone-scale based modulation spectral features (GMT-MSF) on the DNN-SVM framework mentioned in Section 5.4. The obtained performance is reported in Table 6 along with other employed comparison techniques. Our implementation of gammatone-spectrogram includes gammatone filter design using python *Spafe* [94] toolkit, followed by application of the filters over the signal spectrogram. The GMT-MSF is computed over the designed gammatone time-frequency representation following the same steps used to compute CQT-MSF (Fig. 3).

From Table 6 we observe that CQT-MSF outperforms GMT-MSF on EmoDB database but performs relatively poor for RAVDESS database. This observation is due to the difference in non-linearity between constant-Q and gammatone scale. Previous studies on speech recognition [95], speaker recognition [96], and SER [46] also show how different non-linearity during frequency wrapping affect recognition performances. Inspired by those studies, we compare the three concerned non-linear scales used in our experiments: mel, constant-Q, and gammatone in Fig. 10 along with the filterbank center frequencies. For better visibility of differences in filter placement, we use 96 frequency bins for every scale in this analysis. The figure shows that constant-Q scale provides highest

low-frequency emphasis (because of binary logarithm) followed by gammatone and mel-scale. Considering the relevance of the low-frequency information for SER [46, 47], the underperformance of constant-Q scale is an interesting observation.

Previous studies indicate that the dataset-dependent non-linearity scale is more appropriate than a general-purpose scale and providing importance to higher energy region helps in performance optimization [70, 95]. Hence, to further investigate the performance gap, we compare the energy spectral density (ESD) averaged across all utterances for both the databases. Fig. 11 shows the respective averaged energy density plots. The energy densities are computed by time averaging the squared feature coefficients, followed by averaging across all utterances. Fig. 11 shows that when compared to EmoDB, the energy density in RAVDESS is more shifted towards higher frequency regions. The constant-Q scale provides greater emphasis at low-frequency bins (refer Fig. 10) but due to high non-linearity (binary logarithm), the resolution reduces drastically as we move towards higher frequencies. Hence, for RAVDESS, the resolution on the frequencies with the larger ESD value is lower in CQT compared to the resolution provided by gammatone filterbank placed in the ERB scale. Thus, for RAVDESS, GMT-MSF better captures the signal energy information and leads to better performance when compared to CQT-MSF.

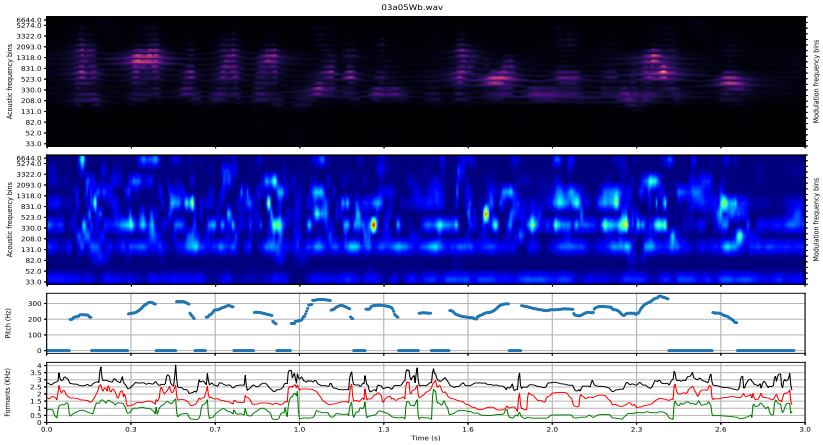
From Table 6 we also observe that the employed EmoNet-based ResNet architecture also shows competitive performance on RAVDESS database when compared to CQT-MSF feature. Larger model size with comparatively larger database leads to this observation.

6.3. Visual Inspection of Learned Features

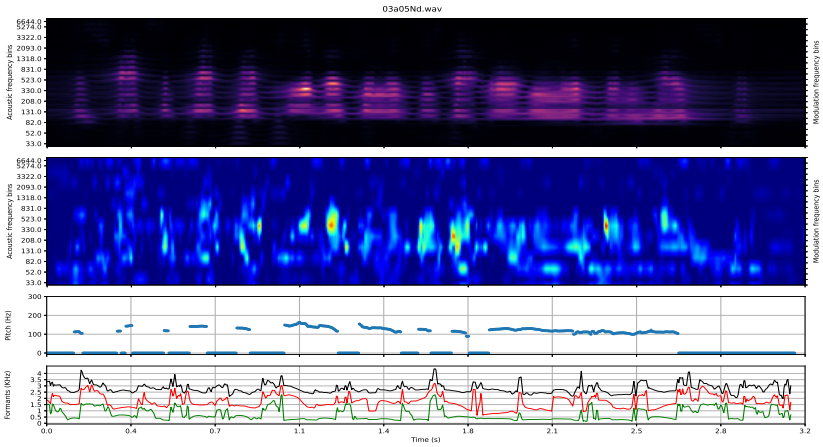
In spite of the performance gain achieved from the proposed CQT-MSF feature, the complexity of the model (Table 1) in terms of network parameters makes it very difficult to understand which information in input helps the network to recognize the pattern. To obtain a general insight into the operation of deep networks, several works use gradient generated at the ultimate layer with respect to the network input to generate a saliency map. This map shows corresponding input regions which weigh the most in generating the output probability scores [97–99]. We use one such analysis, called the *gradient-based class activation mapping* (Grad-CAM), to obtain insight into the working of the model employed [99].

Grad-CAM uses the class-wise gradient generated at network output with respect to the activations of the final convolution layer to generate a *heatmap* showing the importance of different regions of the input. The steps included in Grad-CAM heatmap generation are:

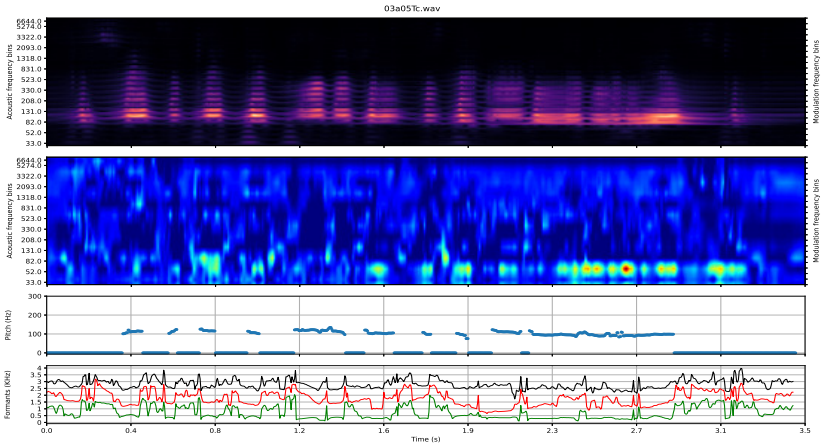
1. Compute the gradient of network output score (before softmax non-linearity) with respect to the activations of final convolution layer, i.e., $\frac{\partial y_c}{\partial A_k}$, where y_c is the score of the cth class and A_k is the 3-dimensional (height, width and channel dimension) class activations from the final convolution layer.
2. Average the computed gradients over length and width dimensions (global average pooling) to obtain a single vector representation of gradients. Mathematically, $\alpha_k = \frac{1}{N} \sum_{\text{length}} \sum_{\text{width}} \frac{\partial y_c}{\partial A_k}$.



(a) Constant-Q MSF with Grad-CAM, pitch and first three formants of *Anger* utterance



(b) Constant-Q MSF with Grad-CAM, pitch and first three formants of *Neutral* utterance



(c) Constant-Q MSF with Grad-CAM, pitch and first three formants of *Sad* utterance

Fig. 12: Grad-CAM output of three different emotion utterances of EmoDB database. To analyse the significance of various frequency regions, pitch and first three formants are also shown along with the Grad-CAM output. For MSF plot, the y-axis labels on the left describe the auditory frequencies and ticks on the right describe the modulation bins corresponding to every auditory frequency bin. The title of every plot describes the EmoDB file name used for analysis.

3. Multiply the computed gradient vector with the final convolution layer activations and average the result over the number of filters in convolution layer, i.e., $map = \frac{1}{N}\sum_{k}(\alpha_k A_k)$.
4. Apply $ReLU$ activation over computed class activation maps in the previous step, $L^{c}_{\text{Grad-CAM}} = ReLU(map)$.
5. Upsample the computed 2-D heatmap over length and width axes to make its shape similar to the input image. The upsampled heatmap shows the importance of various regions of input image which led to the final class prediction.

Fig. 12 shows the Grad-CAM output of the CQT-MSF feature of *Anger*, *Neutral*, and *Sad* utterances of the same speaker (speaker 03) and context (a05) from EmoDB database. For analysis of Grad-CAM output, we also plot the pitch and first three formant frequencies (F1, F2 & F3) of utterances to compare the frequency regions which are most focused upon by the network to predict emotion classes. We observe that pitch and the first two formants (F1 & F2) are important for the emotion class prediction. Lower pitch harmonics are apparently more significant for *Sad* emotion as shown in Fig. 12c. Another important observation for *Sad* is the presence of high Grad-CAM score at the silence and unvoiced regions (where pitch frequency is 0 Hz) of utterance. This shows that the silence between spoken phonemes is also important for emotion recognition. In *Anger* and *Neutral* emotions, formants F1 & F2 are more prominent than pitch. Both emotions are identified mostly in the voiced region of utterances. Interestingly, the Grad-CAM response of *Sad* also shows some focus on high frequencies (near formant F3) over complete utterance as compared to *Neutral* emotion class. Observations made from the Grad-CAM analysis indicate the importance of low frequencies for emotion recognition, especially for low-arousal emotions like *Sad*, further justifying the use of CQT time-frequency representation for SER. Also, MSF representation provides the Conv2D classifier with different modulation rates of different speech characteristics, such as pitch harmonics, formants etc. This helps the classifier to emphasize the emotion-wise differences appearing across different modulation rates, for different speech characteristics (pitch, formant, etc.), hence improving the performance.

6.4. Discussion

Our performed experiments show higher relevance of CQT-based features, as compared to mel-scale features, for SER. We summarise and interpret the results obtained from the experiments in the following points:

- **The two-staged auditory processing based features, i.e., early auditory with cortical analysis based features improve emotion classification performance**. This also justifies the combination of human auditory analysis (domain knowledge) with neural networks for the betterment of SER. However, such improvement is observed over temporal modulations extracted from CQT representation only. Temporal modulations of MFSC show opposite effect and degrade the performance.

- **The improvement observed with CQT-MSF (both fusion types) and standalone CQT over MFSC features show the relevance of increased low-frequency resolution in CQT for SER**. As low-frequency resolution in

mel-scale is not as high, it does not provide enough emphasis over the emotion relevant low-frequencies to capture the information required for emotion discrimination. Hence, its modulation spectrum coefficients also end up with less emotion relevant parts of speech, e.g., irrelevant high frequency regions, leading to reduction in performance.

- **DNN framework lags behind in performance as compared to DNN-SVM classification framework with RBF kernel function**. This observation is consistent across every employed feature. The advantage in performance of DNN-SVM framework has also been mentioned in SER [12] and speaker recognition [82] works.

- **The confusion matrices of different features show a general misclassification trend in *Happy-Anger* and *Fear-Happy* emotion pairs**. This confusion mainly appears due to very similar arousal characteristics of features. However, *Happy-Anger* pair are placed very distant in valence plane. This is due to higher focus on speech prosody in constant-Q representation, and higher sensitivity of speech prosody over arousal characteristics [88]. Although, inclusion of modulations of constant-Q representation reduces the confusion among emotions with opposite valence characteristics.

- **Both CQT-MSF and standalone CQT also outperforms scattering transform coefficients**. Scattering transforms apply averaging over features with empirically defined averaging scale to obtain a translation invariant representation. The convolutional neural network, when used with constant-Q features as input, automatically learns this requires invariance with cross-entropy objective function. Hence, the joint effect of CQT/CQT-MSF and automatic translation invariance makes our frameworks superior. Although, scattering transform manages to outperform mel-scale features, again because of the constant-Q filter banks and time shift and deformation stability in scattering coefficients.

- **CQT-MSF feature outperforms GMT-MSF on EmoDB but underperforms on RAVDESS database.** Comparative analysis shows a slight high-frequency shift in average energy spectral density of the RAVDESS database compared to EmoDB database. The difference in non-linearities of the gammatone and CQT scales result in gammatone-spectrogram better capturing energies with a slight high-frequency shift. This explains the observed anomaly in the performance with RAVDESS database.

Studies in psychology report that individuals with music expertise are better capable of perceiving emotions from speech [55–60, 100, 101]. This finding falls in line with our experiments. As CQT was originally invented for music analysis, its better suitability for SER can be considered as a mathematical evidence, supporting the findings in psychology domain. Another justification towards increased SER suitability of CQT, can be proposed by analysing studies performed over *amusia* in [54, 60]. Amusia is a medical condition in which individuals have limited capability to perceive or resolve pitch. The study in [54] reports that emotion recognition ability of amusic individuals is below par with that of normal individuals, which is attributed to their limited ability to resolve pitch or

low-frequencies of speech. Amusics then utilize the high-frequency content of speech to decipher emotions but are not as efficient as healthy individuals. Therefore, an amusic brain can be assumed to represent speech as a time-frequency representation with low resolution at low-frequencies and comparatively higher resolution at high frequencies. In a contrastive manner, a representation with high low-frequency resolution should improve SER ability, which is what we observe in our experiments with CQT-based features. Also, modulation coefficients computed over CQT is analogous to cortex-level analysis performed over music trained brain. Study performed in [60] reports that such cortical analysis has further beneficial effects over SER ability.

7. Conclusion

This paper proposed the use of constant-Q transform based modulation spectral features for SER. The proposed feature employs the knowledge of two-staged sound processing in humans as domain knowledge and is tested over two different deep network based classification frameworks. We show that the proposed feature outperforms standard mel-frequency based feature and scattering transform coefficients. From the performed experiments, we conclude the following:

1. A representation with increased low-frequency resolution is a better contender for SER. Similar conclusion is endorsed in psychology based studies as well.
2. The combination of a time-frequency representation with higher low-frequency resolution, and its temporal modulation (two-staged representation), efficiently represent the emotion contents in speech.
3. Mel-scale based feature and its temporal modulations are not very significant from speech emotion information perspective.
4. Similar, to mel-scale features, CQT and its temporal modulation representation are also time deformation invariant. With CQT-MSF as input, the CNN can learn the required invariance to time-frequency shifts, leading to a representation stable to shifts and deformations.
5. The DNN-SVM framework provides better SER performance as compared to the standard DNN framework.
6. Grad-CAM based analysis performed over MSF reestablishes the importance of pitch and formant frequencies for SER. It also describes the importance of different modulation rates of pitch and formants, apart from their crude values, for SER.

Although the proposed feature performs better than standard features, the performance is still not optimum for practical real-world deployment of the feature. Also, the performance varies by a large margin when the feature is used over different databases. Even though found more efficient than mel-scale, constant-Q scale is effective for utterances with larger average energy in low-frequency regions. This opens up the opportunity to explore database-dependent non-linear scale for SER. In future, we would also like to experiment with joint spectral and temporal modulation feature and analyse its suitability for SER. To combine the domain knowledge with self-learning, a deep network based architecture for self-learned modulation feature extraction can also be explored.

References

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

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

[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] M. Shah Fahad, A. Ranjan, J. Yadav, A. Deepak, A survey of speech emotion recognition in natural environment, Digital Signal Processing 110 (2021) 102951.

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

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

[7] 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.

[8] 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.

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

[10] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, J. D. Newman, Stress and emotion classification using jitter and shimmer features, in: Proc. ICASSP, volume 4, 2007, pp. IV–1081.

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

[12] 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.

[13] 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.

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

[15] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks, Biomedical Signal Processing and Control 47 (2019) 312–323.

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

[17] 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.

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

[19] 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.

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

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

[22] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[23] M. Kimura, M. Tanaka, New perspective of interpretability of deep neural networks, in: Proc. 3rd International Conference on Information and Computer Technologies (ICICT), IEEE, 2020, pp. 78–85.

[24] S. Shamma, Encoding sound timbre in the auditory system, IETE Journal of Research 49 (2003) 145–156.

[25] N. Muralidhar, M. R. Islam, M. Marwah, A. Karpatne, N. Ramakrishnan, Incorporating prior

domain knowledge into deep neural networks, in: Proc. International Conference on Big Data, 2018.
[26] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, J. Schuecker, Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems, IEEE Transactions on Knowledge and Data Engineering (2021) 1–1.
[27] 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.
[28] M. Goudbeek, J. P. Goldman, K. R. Scherer, Emotion dimensions and formant position, in: Proc. INTERPSEECH, 2009, pp. 1575–1578.
[29] 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.
[30] 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.
[31] T. Chi, P. Ru, S. A. Shamma, Multiresolution spectrotemporal analysis of complex sounds, The Journal of the Acoustical Society of America 118 (2005) 887–906.
[32] S. Kumar, K. von Kriegstein, K. Friston, T. D. Griffiths, Features versus feelings: dissociable representations of the acoustic features and valence of aversive sounds, Journal of Neuroscience 32 (2012) 14184–14192.
[33] L. Arnal, A. Flinker, A. Kleinschmidt, A.-L. Giraud, D. Poeppel, Human screams occupy a privileged niche in the communication soundscape, Current Biology 25 (2015) 2051–2056.
[34] S. V. Vuuren, H. Hermansky, On the importance of components of the modulation spectrum for speaker verification, in: Proc. ICSLP, 1998.
[35] N. H. Sephus, A. D. Lanterman, D. V. Anderson, Modulation spectral features: In pursuit of invariant representations of music with application to unsupervised source identification, Journal of New Music Research 44 (2015) 58–70.
[36] H. Hermansky, History of modulation spectrum in ASR, in: Proc. ICASSP, 2010, pp. 5458–5461.
[37] S. Wu, T. H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Communication 53 (2011) 768–785.
[38] M. J. Alam, Y. Attabi, P. Dumouchel, P. Kenny, D. D. O'Shaughnessy, Amplitude modulation features for emotion recognition from speech., in: Proc. INTERSPEECH, 2013, pp. 2420–2424.
[39] Z. Zhu, R. Miyauchi, Y. Araki, M. Unoki, Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants, in: Proc. INTERSPEECH, 2016, pp. 262–266.
[40] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, M. Akagi, Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends, IEEE Access 8 (2020) 16560–16572.
[41] A. R. Avila, Z. Akhtar, J. F. Santos, D. O'Shaughnessy, T. H. Falk, Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild, IEEE Transactions on Affective Computing 12 (2021) 177–188.
[42] S. R. Kshirsagar, T. H. Falk, Quality-aware bag of modulation spectrum features for robust speech emotion recognition, IEEE Transactions on Affective Computing (2022) 1–14.
[43] A. R. Avila, S. R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, T. H. Falk, Speech-based stress classification based on modulation spectral features and convolutional neural networks, in: Proc. EUSIPCO, 2019, pp. 1–5.
[44] L.-Y. Yeh, T.-S. Chi, Spectro-temporal modulations for robust speech emotion recognition, in: Proc. INTERSPEECH 2010, 2010, pp. 789–792.
[45] Z. Peng, J. Dang, M. Unoki, M. Akagi, Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech, Neural Networks 140 (2021) 261–273.
[46] P. Singh, S. Waldekar, M. Sahidullah, G. Saha, Analysis of constant-Q filterbank based representations for speech emotion recognition, Digital Signal Processing (2022) 103712.
[47] 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, pp. 1–6.
[48] 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.

[49] S. Sukittanon, L. E. Atlas, Modulation frequency features for audio fingerprinting, in: Proc. ICASSP, volume 2, 2002, pp. II–1773–II–1776.
[50] S. Sukittanon, L. Atlas, J. Pitton, Modulation-scale analysis for content identification, IEEE Transactions on Signal Processing 52 (2004) 3023–3035.
[51] H. Hermansky, Speech recognition from spectral dynamics, Sadhana 36 (2011) 729–744.
[52] D. Zotkin, S. Shamma, P. Ru, R. Duraiswami, L. Davis, Pitch and timbre manipulations using cortical representation of sound, in: Proc. ICASSP, volume 5, 2003, pp. V–517.
[53] J. Andén, S. Mallat, Deep scattering spectrum, IEEE Transactions on Signal Processing 62 (2014) 4114–4128.
[54] S. Lolli, A. Lewenstein, J. Basurto, S. Winnik, P. Loui, Sound frequency affects speech emotion perception: results from congenital amusia, Frontiers in Psychology 6 (2015) 1340.
[55] E. Dmitrieva, V. Y. Gel'man, K. Zaitseva, A. Orlov, Ontogenetic features of the psychophysiological mechanisms of perception of the emotional component of speech in musically gifted children, Neuroscience and Behavioral Physiology 36 (2006) 53–62.
[56] C. D. Fuller, J. J. Galvin III, B. Maat, R. H. Free, D. Başkent, The musician effect: Does it persist under degraded pitch conditions of cochlear implant simulations?, Frontiers in neuroscience 8 (2014) 179.
[57] J. T. Twaite, Examining relationships between basic emotion perception and musical training in the prosodic, facial, and lexical channels of communication and in music, City University of New York, 2016.
[58] J. Weijkamp, M. Sadakata, Attention to affective audio-visual information: Comparison between musicians and non-musicians, Psychology of Music 45 (2017) 204–215.
[59] W. F. Thompson, E. G. Schellenberg, G. Husain, Decoding speech prosody: Do music lessons help?, Emotion 4 (2004) 46.
[60] C. Nussbaum, S. R. Schweinberger, Links between musicality and vocal emotion perception, Emotion Review 13 (2021) 211–224.
[61] M. Todisco, H. Delgado, N. Evans, Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification, Computer Speech & Language 45 (2017) 516–535.
[62] C. Schörkhuber, A. Klapuri, Constant-Q transform toolbox for music processing, in: Proc. 7th Sound and Music Computing Conference, 2010, pp. 3–64.
[63] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, T. Kim, Thassilo, librosa/librosa: 0.8.1rc2, 2021.
[64] S. Greenberg, B. Kingsbury, The modulation spectrogram: in pursuit of an invariant representation of speech, in: Proc. ICASSP, volume 3, 1997, pp. 1647–1650 vol.3.
[65] M. Elhilali, Modulation representations for speech and music, in: Timbre: Acoustics, perception, and cognition, Springer, 2019, pp. 335–359.
[66] N. Moritz, J. Anemüller, B. Kollmeier, Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments, in: Proc. ICASSP, 2011, pp. 5492–5495.
[67] R. Banse, K. R. Scherer, Acoustic profiles in vocal emotion expression, Journal of Personality and Social Psychology 70 (1996) 614.
[68] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, D. Poeppel, Temporal modulations in speech and music, Neuroscience & Biobehavioral Reviews 81 (2017) 181–187. The Biology of Language.
[69] 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.
[70] D. Paul, M. Pal, G. Saha, Spectral features for synthetic speech detection, IEEE Journal of Selected Topics in Signal Processing 11 (2017) 605–617.
[71] W. Ghezaiel, L. Brun, O. Lézoray, Hybrid network for end-to-end text-independent speaker identification, in: Proc. International Conference on Pattern Recognition, Milan (virtual), Italy, 2021.
[72] C. Baugé, M. Lagrange, J. Andén, S. Mallat, Representing environmental sounds using the separable scattering transform, in: Proc. ICASSP, 2013, pp. 8667–8671.
[73] P. Singh, G. Saha, M. Sahidullah, Deep scattering network for speech emotion recognition, in: Proc. EUSIPCO, 2021, pp. 131–135.
[74] T. Wiatowski, H. Bölcskei, A mathematical theory of deep convolutional neural networks for feature extraction, IEEE Transactions on Information Theory 64 (2018) 1845–1866.
[75] P. Grohs, T. Wiatowski, H. Bölcskei, Deep convolutional neural networks on cartoon functions,

in: Proc. International Symposium on Information Theory (ISIT), 2016, pp. 1163–1167.
[76] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, A database of German emotional speech, in: Proc. INTERSPEECH, 2005, pp. 1517–1520.
[77] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech Communication 52 (2010) 613–625.
[78] 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.
[79] 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.
[80] S. Ntalampiras, N. Fakotakis, Modeling the temporal evolution of acoustic parameters for speech emotion recognition, IEEE Transactions on Affective Computing 3 (2012) 116–125.
[81] 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).
[82] 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.
[83] 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.
[84] 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.
[85] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011).
[86] A. Rosenberg, Classifying skewed data: Importance weighting to optimize average recall, in: Proc. INTERSPEECH, 2012, pp. 2242–2245.
[87] S. Deb, S. Dandapat, Emotion classification using segmentation of vowel-like and non-vowel-like regions, IEEE Transactions on Affective Computing 10 (2019) 360–373.
[88] I. Luengo, E. Navas, I. Hernáez, Feature analysis and evaluation for automatic emotion identification in speech, IEEE Transactions on Multimedia 12 (2010) 490–501.
[89] A. Aftab, A. Morsali, S. Ghaemmaghami, B. Champagne, LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition, in: Proc. ICASSP, 2022, pp. 6912–6916.
[90] Y. Liu, H. Sun, W. Guan, Y. Xia, Z. Zhao, Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework, Speech Communication 139 (2022) 1–9.
[91] M. Gerczuk, S. Amiriparian, S. Ottl, B. W. Schuller, EmoNet: A transfer learning framework for multi-corpus speech emotion recognition, IEEE Transactions on Affective Computing (2021) 1–1.
[92] E. Guizzo, T. Weyde, J. B. Leveson, Multi-time-scale convolution for emotion recognition from speech audio signals, in: Proc. ICASSP, IEEE, 2020, pp. 6489–6493.
[93] L. F. Parra-Gallego, J. R. Orozco-Arroyave, Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments, Digital Signal Processing 120 (2022) 103286.
[94] A. Malek, S. Borzì, C. H. Nielsen, Superkogito/spafe: v0.2.0, 2022. URL: `https://doi.org/10.5281/zenodo.6824667`. doi:`10.5281/zenodo.6824667`.
[95] K. Paliwal, B. Shannon, J. Lyons, K. Wojcicki, Speech-signal-based frequency warping, IEEE Signal Processing Letters 16 (2009) 319–322.
[96] S. Sarangi, M. Sahidullah, G. Saha, Optimization of data-driven filterbank for automatic speaker verification, Digital Signal Processing 104 (2020) 102795.
[97] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Proc. ICLR, 2014.
[98] J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: Proc. ICLR, 2015.
[99] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proc. International Conference on Computer Vision (ICCV), 2017, pp. 618–626.
[100] 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.
[101] 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.