IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 2, 2020, accepted April 15, 2020, date of publication April 20, 2020, date of current version May 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988781

Emotional Voice Conversion Using a Hybrid Framework With Speaker-Adaptive DNN and Particle-Swarm-Optimized Neural Network

SUSMITHA VEKKOT[1], (Member, IEEE), DEEPA GUPTA[2], MOHAMMED ZAKARIAH[3], AND YOUSEF AJAMI ALOTAIBI[4], (Senior Member, IEEE)

[1]Department of Electronics & Communication Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India
[2]Department of Computer Science & Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India
[3]Research Center, College of Computer and Information Science, King Saud University, Riyadh 11451, Saudi Arabia
[4]Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia

Corresponding author: Deepa Gupta (g_deepa@blr.amrita.edu)

This work was supported in part by the Research Groups Program (Research Group) under Grant RG-1439-033 through the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia, and in part by the Government of India's Visveswaraya Ph.D. scheme through the scholarship for Susmitha Vekkot.

ABSTRACT We propose a hybrid network-based learning framework for speaker-adaptive vocal emotion conversion, tested on three different datasets (languages), namely, EmoDB (German), IITKGP (Telugu), and SAVEE (English). The optimized learning model introduced is unique because of its ability to synthesize emotional speech with an acceptable perceptive quality while preserving speaker characteristics. The multilingual model is extremely beneficial in scenarios wherein emotional training data from a specific target speaker are sparsely available. The proposed model uses speaker-normalized mel-generalized cepstral coefficients for spectral training with data adaptation using the seed data from the target speaker. The fundamental frequency (F0) is transformed using a wavelet synchrosqueezed transform prior to mapping to obtain a sharpened time–frequency representation. Moreover, a feedforward artificial neural network, together with particle swarm optimization, was used for F0 training. Additionally, static-intensity modification was also performed for each test utterance. Using the framework, we were able to capture the spectral and pitch contour variabilities of emotional expression better than with other state-of-the-art methods used in this study. Considering the overall performance scores across datasets, an average melcepstral distortion (MCD) of 4.98 and root mean square error (RMSE-F0) of 10.67 were obtained in objective evaluations, and an average comparative mean opinion score (CMOS) of 3.57 and speaker similarity score of 3.70 were obtained for the proposed framework. Particularly, the best MCD of 4.09 (EmoDB-happiness) and RMSE-F0 of 9.00 (EmoDB-anger) were obtained, along with the maximum CMOS of 3.7 and speaker similarity of 4.6, thereby highlighting the effectiveness of the hybrid network model.

INDEX TERMS ANN, CMOS, DNN, emotion, MCD, MGCEP, PSO, RMSE-F0, speaker-adaptation, speaker similarity score, WSST.

I. INTRODUCTION

Emotions form a salient aspect of human communication via various modalities. However, speech is the most easily accessible data, as it contains various cues such as linguistic information, gender and identity of the speaker, and emotion. Emotions are critical for achieving active dialogue delivery to maintain efficient socio–cultural relationships and enable better human–machine interaction. Furthermore, emotion/affect synthesis can be applied in various domains such as storytelling, speech assistance for the disabled [1]–[3], emotion recognition [4]–[6] and speech-to-speech (S2S) translations [7].

A major application of expression-synthesis models is to provide the speech output of text-to-speech synthesis (TTS) systems that have the required modulation and naturalness. Notably, conversational dialogues frequently involve instinctive and involuntary code switching between

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

languages. Particularly, in multilingual societies such as India, where languages with several dialects co-exist [8], it is challenging to develop a unified dialogue system for S2S translation. Moreover, prosody and spectral-mapping frameworks trained and tested on multiple languages are, especially, useful for designing affective S2S systems in under-resourced languages. Here, the insufficiency of training data can be alleviated by using unified training models, and by performing further testing on sparse languages. Speech emotion can be interpreted from the linguistic data, prosody, and voice quality [9] of an utterance. The parameters used to categorize emotional expressions are based on the spectral features analyzed at the segmental level and prosodic features at the supra-segmental level. Therefore, an emotion-synthesis framework, essentially, modifies the spectral and prosodic components extracted from the neutral speech to those of the target emotion via parameter learning, generally using aligned parallel data.

Predominantly, the research on the synthesis of expressive speech has been focused on source-to-target spectral, or prosodic mapping, or simultaneous modification of both. Previous approaches were focused on rule-based [10], diphone concatenation [11] or signal processing techniques [12]–[19]. In most of these approaches, the feature-transformation scales for suprasegmental prosody were derived from monolingual datasets. Although rule-based approaches are relatively simple and straightforward, the voice quality and naturalism of emotional expression is further enhanced using statistical-modelling approaches. Most experimental frameworks for spectral mapping in emotional-voice conversion used Gaussian mixture models (GMM) based spectral mapping [20]–[26]. Nonetheless, statistical-averaging models such as GMM frequently result in both the loss of important spectral details and spectral over-smoothing. Notably, over-smoothing can be reduced via dynamic feature fusion, incorporation of global variance [27] into the GMM [28], [29] and the application of partial least squares [30].

Modern emotion-conversion schemes are generally data driven, whereas speech-based features must be trained using machine-learning approaches. Non-linear relationships between neutral and emotional features are generated using artificial neural networks (ANNs) and deep neural networks (DNNs) [31]–[35]. Spectral modelling uses mostly non-linear MGCEPs. Recent advances include unsupervised training with conditional restricted Boltzmann machine (CRBM) [36], pre-training using deep belief networks (DBNs) [34], modelling the spectrum and prosody simultaneously via bidirectional long short-term memory (LSTM) [37], end-to-end emotional-speech synthesis using Tacotron [38], among others.

The supra-segmental elements of speech prosody, such as fundamental frequency (F0), speech rate, and energy, have been modelled for expressive-speech synthesis. Among these, pitch or F0 mapping is of utmost interest for representing emotional speech, as even the slightest deviations in prosodic patterns are efficiently captured using the F0 contour trajectory. Pitch mapping is, generally, performed around glottal activity regions, which are determined on the basis of epochs, i.e., the regions of significant excitation of the vocal tract. Because prosody varies non-uniformly across an entire utterance, both static and dynamic prosody transformations [13], [15], [39]–[42] were tested in monolingual datasets. However, most of these approaches model the source and vocal tract independently, thereby failing to capture the correlation between them. Consequently, the converted speech is frequently distorted.

The F0 features for training are generally extracted using the STRAIGHT vocoder [43]. However, these features being lower dimensional, cannot be adequately converted using a DNN [32] or DBN [34]. Therefore, F0 features are frequently converted using linear-transformation-based methods, such as the logarithm Gaussian (LG) method [44]. It has previously been proved that prosody conversion is enhanced upon analyzing the sequence information for modelling short- and long-term dependencies in an utterance [45]. The continuous wavelet transform (CWT) based decomposition of F0 could model the prosody more satisfactorily than the LG baseline in various temporal scales [46]–[48]. A flexible CWT modelling scheme, which uses adaptive scales for various levels of hierarchical prosody, i.e., sentence, phrase, word, and syllable, was proposed in [49], [50].

Only a few adaptation strategies have been tested extensively by emotion-synthesis researchers. Most of these strategies were based on a hidden Markov model using a constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation [51] or training a speaker-dependent model by adapting from speaker-independent average models [52]. In addition, the method in [52] used a segment of duration as short as 5 min of the emotional data from the target speaker and could produce an appreciable perception of the synthesized emotion, however compromising on the speech quality. CSMAPLR considers the linguistic information from the regression tree, thereby distinguishing it from other adaptation approaches. The continuous control of emotional intensity was accomplished by using the combination of a three-layer adaptive neuro-fuzzy inference system and a Fujisaki model for extracting the F0 contour [53]; the method was tested for monolingual data in the Japanese language.

Modern techniques applied in emotion conversion include unsupervised style transfer using generative adversarial networks (GANs) [54], i-vector probabilistic linear discriminant analysis based emotion conversion with non-parallel training [55], sequence-to-sequence F0 modelling and conversion using linguistic conditioning on syllable position [56], and cross-wavelet transform based methods for F0 modelling using combined variational autoencoder GANs [57]. Recently, efforts have been invested in integrating an emotion-conversion module into the convolutional neural network based TTS to improve its naturalism. All these models are data intensive and introduce linguistic

information combined with emotional prosody to simulate affect.

The current emotion-conversion frameworks use learning mainly from multiple aspects that are speaker- or text-dependent. Among the current methods for vocal-emotion conversion, monolingual training is frequently performed from a single-speaker perspective. However, the learning attained from such monolingual approaches is not suitable for applications in multi-speaker or multilingual scenarios. Most modelling techniques are data intensive, and, therefore, several hours of training are required to devise an appropriate transformation function for the selected feature. The adaptation techniques presented in the literature frequently utilize the transplant of emotions to a neutral average model. Frequently, linguistic conditioning is used to attain a reasonable perceptive quality of the converted speech. The unavailability of a large amount of labelled emotional data, need for complex computations, and lack of mature algorithms for sparse-language feature modelling are the main challenges encountered in emotion-synthesis systems.

This study aims to address some of the aforementioned challenges by proposing an adaptation model for emotion conversion; the model was tested on three languages: German, Telugu, and English. During the experimentation, instead of achieving technical prowess, we primarily aimed at synthesizing basic emotions with an acceptable perceptive quality by using as less training data as possible. Accordingly, a framework was established for spectral, F0, and intensity modelling in multiple languages by using less training data and computational overheads. Spectral mapping for MGCEPs was executed using a DNN with speaker adaptation. The framework, which used sparse training data from the target speaker for mapping, achieved appreciable perceptive quality of the synthesized speech. Because the extracted F0 features are discrete and single dimensional, it is challenging to model the temporal inflexions in F0. Wavelet synchrosqueezed transform (WSST) decomposition of F0 is utilized in our work as it bypasses the uncertainty in time-frequency representation by reassigning coefficients to provide a sharpened and complete representation. Multi-layer ANN modelling with particle-swarm-optimized weights and biases was performed for WSST-F0 mapping. To the best of our knowledge, the usage of improved resolution wavelet synchrosqueezing and PSO–ANN (PSO: particle swarm optimization) for F0 modelling in this study is the first of its kind in emotion-conversion approaches.

The following are the major contributions of this work:

- An MLP DNN-based speech-emotion-conversion model with speaker-adaptive training for MGCEP mapping was designed and implemented
- The WSST decomposition for F0 was implemented and subjected to dimension reduction using principal component analysis (PCA)
- The PCA-reduced WSST-F0 was transformed using a particle-swarm-optimized multilayer feedforward ANN
- The intensity of the converted speech samples was modified to improve the perception of the expression
- The performance of the proposed framework was evaluated both objectively and subjectively
- The efficacy of the proposed model was compared with that of the state-of-the-art methods

The study is organized as follows. Section II details the mathematical theory on which the framework is based, while Section III describes the high-level architecture of the proposed framework. The standard datasets used in this study are provided in Section IV. Furthermore, Section V provides the implementation details of the proposed framework, and Section VI details the evaluation criteria and comparison with other methods. Section VII discusses the results, and Section VIII presents the conclusion with insights for possible extension of the work.

The following section describes the mathematical theory applied in the proposed framework.

II. MATHEMATICAL BACKGROUND

The proposed method uses a hybrid network model with two kinds of neural network architectures for parameter mapping viz. multi-layer perceptron (MLP) DNN mapping for spectral MGCEP and multi-layer ANN with PSO for F0 mapping. Multi-layer perceptron (MLP) DNNs are typical deep-learning models, and they are used for function approximation, with a deeper architecture and regularization compared to ANN. The major decisions involved in selecting an appropriate deep-learning configuration for parameter mapping are with respect to the network architecture, i.e., the number of layers, type of connections, and neurons per hidden layer. The technical details regarding the implementation of DNNs are described in Section. V. For basic theory regarding ANNs and MLP DNNs, kindly refer to [58]–[63].

For a convenient and crisp representation, the MLP DNN structure used in the context of this work is designated simply as 'DNN' throughout the paper. The proposed model further uses WSST decomposition followed by mapping using PSO-ANN for F0 mapping. The following subsections describe the mathematical background behind WSST and PSO.

A. WAVELET SYNCHROSQUEEZED TRANSFORM

Synchrosqueezing is a term used in auditory analysis. It was devised for the decomposition of signals with time-varying characteristics. It belongs to the category of time–frequency reassignment (TFR) algorithms. Compared with classical TFR techniques, synchrosqueezing efficiently reconstructs the components that constitute the time-domain signal. Thus, synchrosqueezing can be used as a substitute to empirical mode decomposition techniques [64]. The concept of WSST has been used for epoch extraction from the emotional speech because of its sharpness in instantaneous frequency representation [65]. In addition, synchrosqueezing successfully represents the rapidly varying pitch in emotional speech, which forms the basis of this study.

Basically, synchrosqueezing aims to acuminate the time-frequency representation given by $R(t, \omega)$ by allotting its value to a distinct point (t_1, ω_1) in the same plane which is ascertained using the local behaviour of $R(t, \omega)$. The synchrosqueezed transform operates on the CWT of a signal (1) [66]

$$W_x(a, b) = \int x(t) a^{-1/2} \Psi(\frac{t-b}{a}) dt \tag{1}$$

where Ψ represents mother wavelet, and a and b represent scale and translation factors of the wavelet, respectively. The instantaneous frequency can be extracted from this signal. According to the observations in [67], although $W_x(a, b)$ is spread-out in a, its oscillatory behaviour is insensitive to the value of a. This means that if we take any wavelet Ψ which is concentrated only on positive axis in frequency, i.e. $\Psi(\xi) = 0$ for all $\xi < 0$, then, by Plancherel's theorem [64],

$$W_x(a, b) = \frac{1}{2\pi} \int x(\xi) a^{0.5} \Psi(a\xi) e^{ib\xi} d\xi \tag{2}$$

If $\Psi(\xi)$ is concentrated around $\xi = \omega_0$, then $W_x(a, b)$ will be around $a = \frac{\omega_0}{\omega}$. For any a and b for which $W_x(a, b) \neq 0$, the instantaneous frequency can be estimated as in (3) [68]:

$$\omega_x(a, b) = -i(W_x(a, b))^{-1} \frac{\partial}{\partial b} W_x(a, b) \tag{3}$$

where ω and a are binned. Therefore,

$$\omega(a, b) = \omega_x(a, b) \tag{4}$$

Practically, because a, b and ω are discrete, $W_x(a, b)$ is computed purely at distinct scales a_k where

$$a_k - a_{k-1} = (\Delta a)_k \tag{5}$$

The synchrosqueezed transform is evaluated only at the center frequencies, ω_c, of consecutive bins, i.e., $[\omega_c - 0.5\Delta\omega, \omega_c + 0.5\Delta\omega]$ where $\Delta\omega = \omega_c - \omega_{c-1}$. The wavelet synchrosqueezed transform can be estimated as in [64]:

$$T_x(\omega_c, b) = (\Delta\omega)^{-1} \sum_{a_k : |\omega(a_k, b) - \omega_c| \leq 0.5\Delta\omega} W_x(a_k, b) a_k^{-3/2} (\Delta a_k) \tag{6}$$

Reconstruction can be performed by taking the inverse transform of T_x. Starting from (7),

$$\int_0^\infty W_x(a, b) a^{-3/2} da = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_0^\infty x(\xi) \overline{\Psi(a\xi)} e^{ib\xi} a^{-1} da d\xi \tag{7}$$

Rearranging the terms and re-writing (7)

$$\int_0^\infty \overline{\Psi(\xi)} \frac{d\xi}{\xi} \frac{1}{2\pi} \int_0^\infty x(\xi) e^{ib\xi} d\xi \tag{8}$$

Considering

$$C_\psi = 0.5 \int_0^\infty \overline{\Psi(\xi)} \frac{d\xi}{\xi} \tag{9}$$

Since x is real,

$$x(\xi) = \overline{x(-\xi)} \tag{10}$$

$$x(b) = \mathrm{Re}\left[\int_0^\infty W_x(a, b) a^{-3/2} da / C_\psi \right] \tag{11}$$

By linear approximation [68],

$$x(b) \approx \mathrm{Re}\left[\sum_c T_x(\omega_c, b) \Delta\omega / C_\psi \right] \tag{12}$$

B. PARTICLE SWARM OPTIMIZATION

PSO aims to make a connection between evolutionary-computation and genetic-algorithm perspectives. Its main attraction is its simplified concept and ease of implementation, requiring no expensive computing capabilities. It is frequently used for optimizing continuous non-linear functions. The idea was derived from natural bird flocking, wherein an optimal distance is ensured among neighbours to avoid collision. Each particle in PSO is initialized with a position and velocity. In addition, the velocity of nearest neighbour is also set to the velocity of the considered particle to create synchronized motion. A stochastic variable is added to the selected velocity at every iteration to create the required variability in the system. Each particle can be regarded as an optimized solution to the problem of determining the flight behaviour [69]. Considering N particles in D dimensions with the optimum position represented as $p_i, i = \{1, 2, \ldots N\}$, the optimal position of an entire population is given as p_g. The velocity v_i and position p_i are dynamically updated by the expressions in (13) and (14) [70]:

$$v_{id}(t+1) = w v_{id}(t) + a_1 r_1 (p_{id}(t) - x_{id}(t)) + a_2 r_2 (p_{gd}(t) - x_{gd}(t)) \tag{13}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \tag{14}$$

where t denotes the number of iterations, $d = \{1, 2, \ldots D\}$, w the inertia weight, i the particle number, a_1, a_2 the acceleration coefficients and r_1, r_2 uniformly distributed random numbers.

The convergence of the algorithm is estimated using the values of the control parameters. The fitness values correlate with the value of the objective function, and they can be used to measure the position and update a particle. Finally, the position of each particle should converge to the optimal particle position of the entire population. The advantage of adopting the PSO algorithm instead of traditional algorithms is that possible solutions will navigate the problem hyperspace quickly toward achieving the global optimum [71]. Therefore, the training provides optimum weights with the lowest possible output deviation. For further details regarding PSO, kindly refer to [71]–[74].

III. PROPOSED FRAMEWORK

The model proposed in this paper has the following main stages:

1) Feature extraction, normalization and dimension reduction
2) Feature mapping
3) Testing and resynthesis

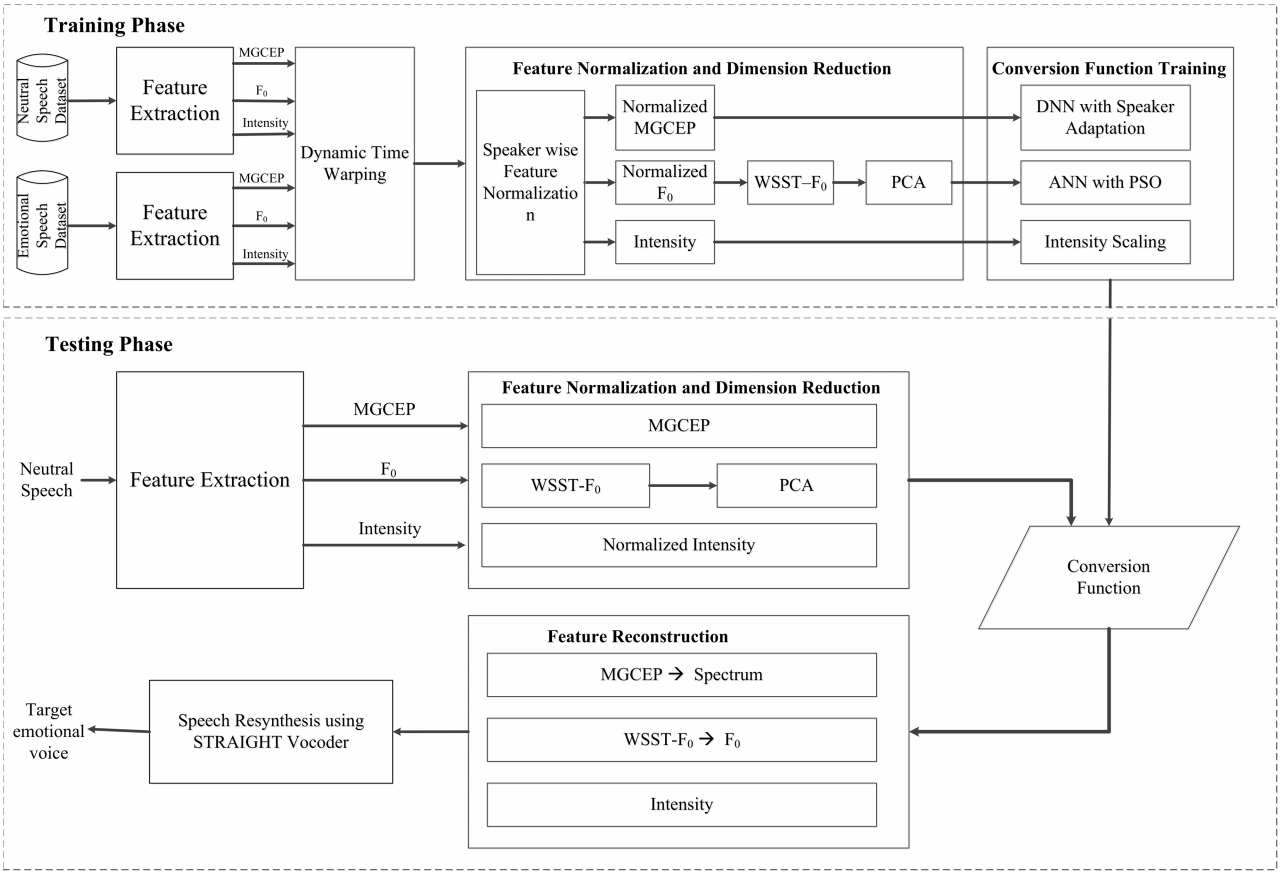


FIGURE 1. High-level architecture of the proposed DNN + WSST-ANN-PSO + Int framework for emotion conversion.

The architecture of the proposed model is depicted in Fig. 1, which is divided into training and testing phases. The training phase involves the extraction, speaker-wise normalization, and mapping strategy devised for the features considered. The testing phase depicted in Fig. 1 contains unseen neutral utterance from the target speaker being passed through conversion function and further feature reconstruction and speech resynthesis. The proposed framework is named the *DNN + WSST-ANN-PSO + Int* model.

Each module is described in the following subsections.

A. FEATURE EXTRACTION, NORMALIZATION AND DIMENSION REDUCTION

The features extracted for processing are MGCEPs, F0, and intensity from parallel neutral and target emotional utterances. MGCEPs are extracted from the spectral representation, and F0 is extracted for every 5 ms. Intensity is extracted at the utterance level. All these features are time-aligned via dynamic time warping before processing. The features are further preprocessed as required to nullify the emotional variabilities. All the features are mean-and-variance normalized speaker-wise to retain the emotional characteristics and reduce the speaker-induced variabilities as much as possible during the emotional utterance by various speakers.

F0 features are decomposed using WSST to yield WSST-F0. Without disturbing the time resolution of the representation, WSST is utilized to reassign the signal energy in the frequency domain by using a phase transform. The time resolution must be preserved to perfectly reconstruct the signal. In addition, the instantaneous frequency information is captured in this representation using an analytic wavelet. Because the resulting WSST-F0 is high dimensional, the dimensionality is reduced using PCA. Therefore, the new WSST-F0 with reduced dimensions is utilized for F0 mapping. Normalized intensities across the entire utterance are used for the mapping.

B. FEATURE MAPPING

The proposed hybrid modelling framework uses a combination of a DNN and an ANN for performing feature mapping and emotional-voice transformation. Spectral mapping is accomplished using an MLP DNN with 10 hidden layers. The aim of performing feature mapping from multiple speakers is to convert the spectral and prosody characteristics from the source to the target without compromising the identifiability of the target speaker. To achieve that aim, an adaptation strategy is used, wherein limited seed data from the target speaker (approximately 1 min) are used for model adaptation

after the first stage of training. The training is repeated using the adaptation data for the features considered.

The PCA-reduced WSST-F0 is fed to a two-layer shallow ANN based on conjugate gradient learning. Instead of opting for straightforward training using the ANN, the weights and biases are further optimized using PSO to yield a simplified and robust conversion function.

Intensity mapping is performed by considering the static intensities across the entire utterance. The scale factor is derived as follows:

$$Scalefactor = \frac{Average\ \ intensity\ \ of\ \ target\ \ emotion}{Average\ \ intensity\ \ of\ \ neutral} \quad (15)$$

C. TESTING AND RE-SYNTHESIS

For testing, the leave-one-speaker-out cross-validation strategy was applied to each dataset. The unseen neutral utterance was provided for the feature-extraction stage to separate the MGCEPs, F0, and intensity. After feature preprocessing, as depicted in Fig. 1, the parameters were fed to the conversion function obtained from the training phase. The conversion-function block in Fig. 1 encompasses separate mapping functions for for MGCEP, WSST-F0 and intensity. The parameters were denormalized and reconstructed before being resynthesized. Notably, after the conversion, the reduced WSST-F0 must be reconstructed to the original dimension prior to performing feature reconstruction. An approximate reconstruction was performed using the PCA principle as follows:

$$WSST - F0_{orig.} = PCA_{red} \times (E)^T + \mu \quad (16)$$

where $WSST - F0_{orig.}$ denotes the original dimension WSST-F0, PCA_{red}, the PCA scores, E, the Eigen vectors and μ represents the mean vector for training data corresponding to the target emotion.

Emotional speech was synthesized using the modified spectral and F0 features by employing the STRAIGHT vocoder. Furthermore, the intensity of the resynthesized utterance was modified using the scale factors obtained in (15) for each dataset.

IV. EXPERIMENTAL DATA SOURCE

To develop an integrated framework for speaker-adaptive vocal-emotion conversion, the selected data must capture the emotional rendering by various speakers. Table 1 lists the datasets used in this work.

The datasets listed in the table were considered because they contain multiple utterances from speakers of both genders for three emotions, namely, anger, fear, and happiness. Furthermore, training and testing in feature-mapping experiments require parallel neutral data in the datasets. In addition, multilingual testing requires the selection of languages that vary with respect to their phonetic content and variabilities in emotional expression. All the recordings in the training dataset and resynthesized emotional speech data were saved in the "*.wav" format as monochannel sound files at the sampling rate of 16 kHz. For the SAVEE dataset [77], the original recording, which was sampled at 44.1 kHz, was resampled to 16 kHz to maintain uniformity throughout the experimentation.

TABLE 1. Experimental datasets.: The total number of speakers, as well as the training and testing utterances across all the datasets, are provided.

Dataset (Language)	No. of Speakers	No. of Training Uttr.	No. of Testing Uttr.
Berlin Emotional Speech Database (EmoDB)(German) [75]	10 Male-5, Female-5	180	20
IIT Kharagpur Simulated Emotion Speech Corpus (IITKGP)(Telugu) [76]	10 Male-5, Female-5	540	60
Surrey Audio-Visual Expressed Emotion (SAVEE)(English) [77]	4 Male-4 Female-Nil	180	20
Total	**24**	**900**	**100**

Although the datasets listed in Table. 1 also contain emotional utterances representing other archetypal emotions, such as disgust, surprise, and boredom, this study primarily focused on the synthesis of the three basic emotions, namely, anger, fear, and happiness. This is because these emotions frequently arise in conversations and can be perceived more satisfactorily in listening tests with converted utterances. In addition, the emotions considered reflect the shifts in the manner of emotional rendering by multiple male and female speakers.

V. EXPERIMENTAL SET-UP

The framework proposed uses three methods of parameter mapping: DNN with speaker adaptation; F0 with WSST-ANN-PSO and static-intensity mapping, respectively, for spectral MGCEPs and F0; and intensity. The experiments were conducted using 1000 utterances from all the datasets, as listed in Table. 1. The feature data were speaker-normalized before being fed to the respective training networks. For performing speaker-wise mean–variance normalization, we normalized the MGCEP and F0 features corresponding to each emotion from each speaker with respect to the mean and variance for the neutral utterance from the same speaker. This resulted in a balanced representation of the feature space data by averaging out speaker-specific traits and retaining only emotion- specific traits. Both male and female speakers were trained separately, and the results projected represent the average from both the genders. The testing was conducted using the leave-one-speaker-out cross-validation strategy. The training was speaker-adaptive and used the seed adaptation data of less than 1 min for the modelling. The mean duration of utterances and the total amount of sound data used from each dataset are provided in Table. 2.

The parameters that must be optimized for spectral training are DNN hyperparameters. For spectral training, 25-dimension MGCEPs were extracted by using the analysis

TABLE 2. Mean Duration of utterance(s) and the amount of speech data used for emotions in each dataset.

Dataset	Emotion	Mean Dur. (s)	Data Used (MB)
EmoDB	Neutral	2.34	25.5
	Anger	2.73	26.7
	Fear	2.24	27.7
	Happiness	2.44	25.0
IITKGP	Neutral	2.05	204.0
	Anger	1.38	182.0
	Fear	2.00	219.0
	Happiness	2.04	205.0
SAVEE	Neutral	3.14	20.5
	Anger	3.17	20.5
	Fear	3.38	22.6
	Happiness	3.54	21.0

frame length of 25 ms and a frame shift of 10 ms. In addition, a DNN with 10 hidden layers was used; it contained 50 neurons per hidden layer. The ReLu activation function was applied to each layer, except the output layer. However, a linear activation function was used at the output layer. An Adam optimizer was used to control the learning-rate parameter in the network. The DNNs were optimized using stochastic gradient descent with MSE as the loss function. The number of epochs was set to 250, with a minibatch size of 30. For regularization, a 20% drop-out was added in the hidden layers. The data were split into training and test sets at the ratio of 80:20. The early stopping criterion was set to 10 epochs; i.e., if the validation MSE did not improve within 10 epochs, the training stopped. In addition, the learning rate was set to 0.002 for the initial 10 epochs and 0.001 for the succeeding epochs with a momentum of 0.9. During speaker adaptation as well, the momentum was set to 0.9 and the learning rate to 0.001.

For F0 mapping from the neutral to the target emotion, the parameters were tuned in the manner described herein. For F0 feature extraction, 25-ms analysis frames with 5-ms frame shift were employed using a robust algorithm for pitch tracking [78], with frequencies varying from 50 to 500 Hz. For wavelet decomposition and ANN mapping, the F0 contour must be continuous. However, because we consider entire utterances for training, the voiced and unvoiced parts must be separated from each other, following which the F0 values are estimated for all the voiced parts in each utterance. Because F0 is zero for the unvoiced speech, the unvoiced portions in the F0 contour are filled via linear interpolation to reduce the discontinuities in the contour while performing ANN mapping. Furthermore, WSST-based transformation and PCA-based dimension reduction were used prior to performing the F0 mapping. WSST was evaluated using the analytic Morlet wavelet at 10 voices/octave. In addition, PCA was applied to the higher-dimensional WSST-F0 to identify the principal components. Because the primary aim of our experiments was to generate the mapping of an acceptable quality by using minimum computational overhead, the dimension was reduced to 12. Moreover, it was observed that a further increase in the dimension did not significantly affect the root mean square error (RMSE). On the basis of the dimension fixed, ANN mapping using a two-hidden-layer neural net with scaled conjugate gradient training was applied to WSST-F0. The number of neurons per hidden layer was empirically set to 30. The *Tanh* activation function was used in both the hidden layers.

F0 mapping was achieved using a multilayer ANN coupled with PSO for the further optimization of the weights and biases. The F0 feature vector was split into training, validation, and test sets in the proportion: 70%, 15%, and 15%, respectively. For the feedforward ANN training, to counter the overfitting to the training data, six validation checks were conducted at every epoch. In addition, MSE was used for scrutinizing the validation error.

The training continued until the validation error decreased continuously. The instance wherein the MSE reached the minimum point of decrease and further started increasing was regarded as the stopping criterion for the training. The best validation performance obtained for each emotion is depicted in Fig. 2. A similar mode of training was conducted for all the datasets, and the epochs corresponding to the best validation was used as the stopping criteria for countering the overfitting.

VI. PERFORMANCE EVALUATION METRICS AND COMPARISON WITH STATE-OF-THE-ART METHODS

To estimate the emotion-conversion efficiency for the datasets, several objective measures were adopted; they are discussed in Subsection. VI-A. Because emotional expression is highly subjective, the objective evaluations were reiterated via subjective measures as well, as described in Subsection. VI-B.

A. OBJECTIVE METRICS

The objective measures used for evaluating the feature-mapping effectiveness are melcepstral distortion (MCD) for spectral-conversion efficiency, root MSE for F0 mapping (RMSE-F0), and perceptual evaluation of speech quality (PESQ) for evaluating the overall quality of the converted speech.

1) MELCEPSTRAL DISTORTION

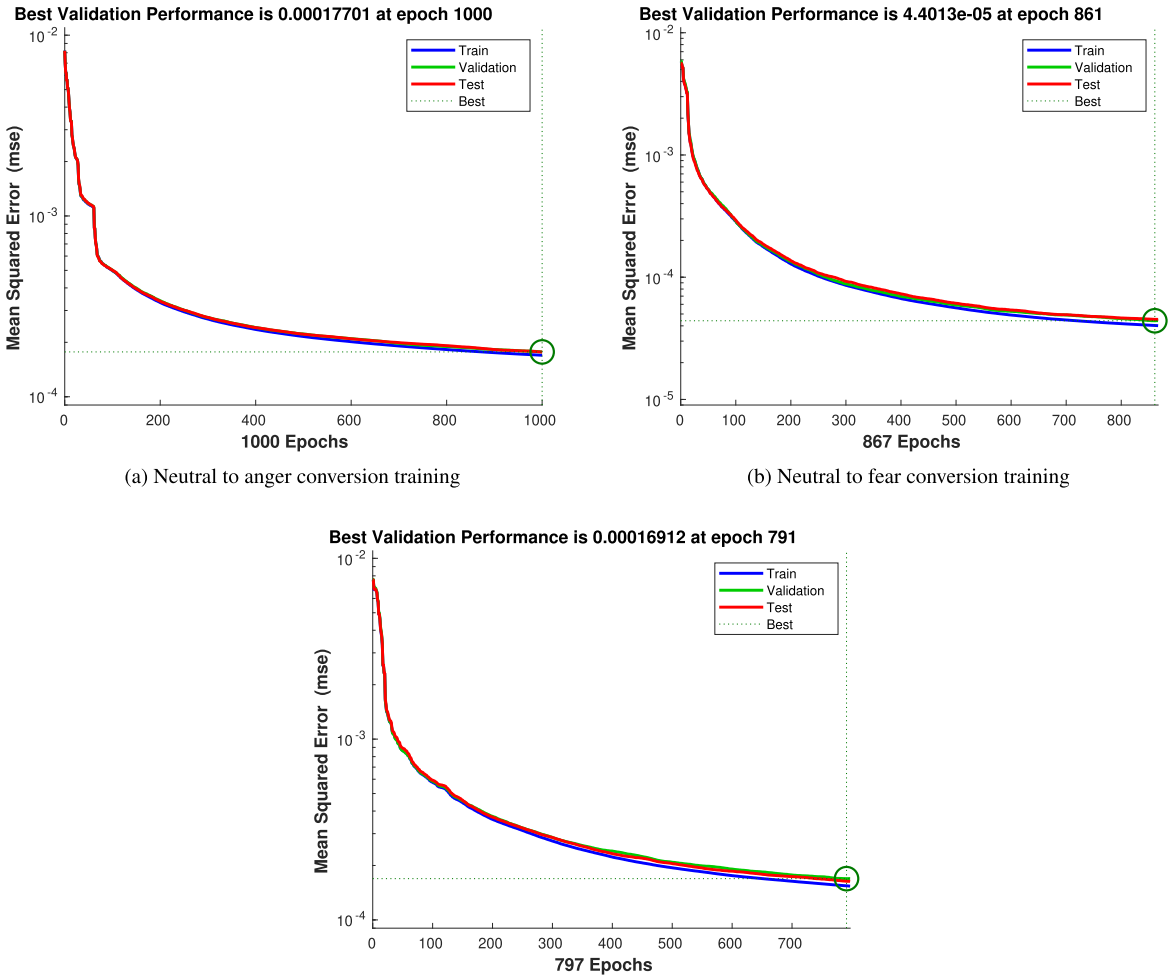
MCD was used to assess the spectral-conversion efficiency [23], [33], [34], [37], [49], [50], [52], [57]. It is calculated as follows

$$MCD = \frac{10}{ln(10)}\sqrt{2\sum_{i=1}^{D}(m_i^t - m_i^c)^2} \tag{17}$$

where m_i^t denotes the i^{th} frame melcepstral coefficient of the target and m_i^c that of the converted utterance; in addition, D denotes the melcepstrum dimension.

2) ROOT MEAN SQUARE ERROR

RMSE is an established metric used [33], [34], [37], [49], [50], [52], [57] to evaluate the proximity of predicted values



(a) Neutral to anger conversion training

(b) Neutral to fear conversion training

(c) Neutral to happiness conversion training

FIGURE 2. Best validation performance obtained and the number of epochs taken to achieve the best MSE for training of WSST-F0 with ANN-PSO. For illustration, the MSE values during training in the EmoDB dataset are plotted in the figure.

by a mapping algorithm to those of the target, given as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(F0_{conv}(i) - F0_{tar}(i))^2}{N}} \tag{18}$$

where $F0_{conv}$ denotes the $F0$ of the re-synthesized emotional utterance and $F0_{tar}$ that of the target emotion. In addition, N denotes the number of samples considered for conversion in each dataset.

3) PERCEPTUAL EVALUATION OF SPEECH QUALITY

PESQ is an objective measure to evaluate the perceptual quality of the enhanced speech. The International Telecommunications Union Standardization Sector has accepted PESQ as recommendation P.862 [79]–[81]. Because the proposed framework generates emotional speech via resynthesis after decomposition using WSST and subsequent PCA, the quality of the converted speech must be maintained post conversion so that the speech is not degraded upon using the mapping algorithm. Moreover, PESQ is used to measure the degradation in the voice quality of the synthesized speech post conversion with respect to the target clean speech in emotional environments [82], [83]. However, the PESQ values reported are MOS-LQO scores calculated to objectively predict the quality of synthesized speech for a listening-only test situation [79]. Thus, the PESQ scores can be directly correlated with that obtained from listening tests.

B. SUBJECTIVE MEASURES

The converted expressive speech is evaluated using objective-evaluation metrics, as discussed in Section VII-A. Nevertheless, as emotion is, to a considerable extent, subjective, the objective-evaluation results must be emphasized further using the aspect of human perception. Accordingly, a subjective evaluation was performed using the comparison mean opinion score (CMOS) of the evaluation measures, and speaker similarity. CMOS/MOS tests have been used for drawing similarities between the synthesized and target emotions [13], [15], [17], [33], [57], [84]–[86], while speaker-similarity scores provide the extent to which the

TABLE 3. Ranking scale used for perception test (CMOS).

Perceptual difference between utterances	CMOS
Identical	5
Almost identical	4
Average similarity	3
slightly different	2
Very different	1

TABLE 4. Ranking scale used for speaker similarity test.

Perceptual similarity	Speaker similarity score
Identical	5
Almost identical	4
Average similarity	3
slightly different	2
Very different	1

identity of a speaker is preserved after conversion. The ranking scales used for estimating CMOS and speaker similarity are explained in Tables. 3 and 4, respectively.

C. COMPARISON WITH STATE-OF-THE-ART METHODS

To analyze the capability of the proposed framework in mapping the three prominent features relevant to emotional expression, it was compared with different variations of the current methods in the literature, termed ''baselines'' in this study. In each comparison, the parameters for the training were the same as those used in the experiments. Intensity mapping was performed using the static utterance-level scale factors provided in (15), whereas one of the other parameter-mapping methods, for either spectrum or F0, was varied. Therefore, the following two categories of mapping methods were used in this study to compare the efficiency of the emotion conversion with that of the proposed *DNN + WSST-ANN-PSO + Int* framework, which are as follows:

- **Spectral mapping baseline**: The baseline used for spectral mapping is DNN without any speaker adaptation. Instead of using seed adaptation data from the target speaker, the speaker-normalized MGCEPs were used for conversion function training. Subsequently, F0 mapping was performed using the proposed WSST-ANN-PSO algorithm. Therefore, this method is termed as *DNN w/o speaker adaptation + WSST-ANN-PSO + Int* hereinafter.

After selecting the spectral-mapping method, we selected several methods to compare their F0-mapping performance with that of the selected method. Here, the spectral mapping was kept fixed as DNN, and different variations were tested in terms of F0 mapping, which are described below. No parameter was optimized for any of the contemporary methods.

- **F0 mapping baselines**: Three variations of state-of-the-art methods were compared with the proposed model in terms of the F0-mapping performance, as illustrated below. In methods 1 and 2, CWT decomposition was used in emotion-conversion experiments, as it facilitates the capturing of the hierarchical information in an utterance at the sentence-, word-, phrase-, and syllable levels [49], [50]. We used the following different variations of mapping methods for performing comparison after CWT decomposition.
  1) **DNN + CWT-NMF + Int**: A DNN with speaker adaptation was used to implement the spectral conversion method; however, for F0 mapping, nonnegative matrix factorization (NMF) was used for the transformation of 30-scale CWT-F0, as in [49].
  2) **DNN + CWT-ANN + Int**: This model is closely related to that proposed in [34]. However, the difference from the method discussed in [34] is that instead of pretraining using a DBN, a DNN with speaker adaptation was used to convert the spectral MGCEP. A feedforward ANN without PSO was utilized to modify the 30-scale CWT-F0.

Although CWT-based methods can capture hierarchical information, WSST-based transformation provides a sharpened F0 representation by means of reduced energy smearing, which is especially useful for spectral reconstruction. To analyze the strength of WSST-F0 irrespective of the mapping technique employed, NMF-based mapping was performed after WSST de-composition instead of using the state-of-the-art CWT. Therefore, a mapping method was created and tested as the following method 3: resume

1) **DNN + WSST-NMF + Int**: This model uses the DNN with speaker adaptation to convert MGCEPs and NMF is used to transform 30-scale WSST-F0 features.

VII. RESULTS AND DISCUSSION

Evaluation of proposed framework alongwith comparison with other feature mapping methods was performed using the performance criteria discussed in Section. VI; details are provided in the following subsections.

A. OBJECTIVE EVALUATION AND ANALYSIS

To analyze the strength of the proposed model for emotion conversion, objective comparisons were performed using standard metrics. The proposed *DNN + WSST-ANN-PSO + Int* framework was compared with all the other methods, discussed in Subsection. VI-C. The first level of comparison was performed using spectral evaluation as the criterion. Here, the proposed model for spectral mapping, i.e., the DNN with speaker adaptation, was compared with that without speaker adaptation using MCD as the performance measure. The results of this comparison are presented in Table. 5. In Table. 5, across all datasets, lower distortion is obtained for DNN mapping with speaker adaptation than without utilizing an adaptation mechanism, using the same method for F0 and intensity modification in both cases. Average MCD for all the emotions considered across datasets was 4.98.

The utilization of seed data for speaker-specific modelling helps preserve target speaker characteristics considerably

TABLE 5. Mel Cepstral Distortion - DNN w/o and with speaker adaptation.

Dataset	Mel Cepstral Distortion (MCD)(dB)					
	DNN w/o spk. adap.+ WSST-ANN-PSO + Int			DNN + WSST-ANN-PSO + Int		
	N2A	N2F	N2H	N2A	N2F	N2H
EmoDB	5.05	4.95	5.24	4.71	4.46	4.09
IITKGP	4.86	5.16	5.04	4.70	4.99	4.76
SAVEE	6.73	7.28	8.02	5.50	5.66	5.99
Average MCD	5.55	5.80	6.10	4.97	5.04	4.95
Performance Improvement (%)						
EmoDB	—	—	—	6.73	9.90	21.95
IITKGP	—	—	—	3.29	3.29	5.56
SAVEE	—	—	—	18.28	22.25	25.31

* N2A:Neutral to Anger, N2F:Neutral to Fear, N2H:Neutral to Happiness

better than a generalized speaker normalization mechanism. Whereas the proposed mapping method performs equally well across all datasets, the performance improvement scores show maximum enhancement in the SAVEE dataset, where the available training data is lower than those in the other two datasets used in the experiments. This decrease in data requirements indicates applications in low resource mapping wherein labeled data accessible from a particular language/speaker is limited.

The efficacy in terms of F0 mapping was compared with that of three state-of-the-art mapping methods: *DNN+CWT-NMF+Int*, *DNN+CWT-ANN+Int* and *DNN+WSST-NMF+Int*. The results obtained are illustrated in Fig. 3. In the figure, the lowest F0-RMSE is obtained for the proposed model across all datasets. An RMSE value as low as nine is obtained for conversion to anger, and conversions to fear and happiness also shows a comparable performance. The average RMSEs obtained by the proposed mapping technique across datasets for each emotion are 10.65 (anger), 11.00 (fear), and 10.37 (happiness). Across datasets, a mean RMSE-F0 of 10.67 was obtained. CWT-based methods yield higher RMSEs than those with WSST, with the CWT-NMF combination giving the highest RMSE in all cases.

Whereas CWT provides a better time-frequency localization ability, the usage of finite duration analysis windows limits the time-frequency resolution and leads to spectral energy smearing, thereby producing artefacts in the representations. By applying the WSST, the readability of the transform is enhanced as instantaneous frequency is computed. Further, the frequency reassignment in the WSST improves the sharpness of the representation. Here, the reduced spectral smearing leads to a crisper time-frequency picture by removing high frequency noise contamination, if any, in the signal. Therefore, a well-defined representation is obtained for each frame in the event of rapid pitch variations, which can enhance mapping for better prediction of F0. The reduced RMSE in both WSST-based F0 mapping cases can be attributed to this.

Albeit both WSST-based methods show a comparable performance in most cases, the optimization strategy applied to the ANN weights further improved the RMSE. The convergence of the ANN-PSO combination was found to be more efficient than that of the ANN alone. In addition, the RMSE generated was lower than that without optimization. The advantage of using PSO is that the convergence to the best solution is guaranteed rather than the algorithm becoming trapped in a not so optimal local solution based on the trial and error strategy used in conventional ANN training.

Although comparable RMSE values are obtained for the proposed framework in the EmoDB and IITKGP datasets, a greater improvement in performance is observed in EmoDB, substantiated by the increased reduction in RMSE (almost to half that of the baseline CWT). In the SAVEE dataset, the optimized algorithm performed better for fear than for anger and happiness. Thus, the optimized pitch mapping algorithm performs better even for unbalanced datasets, such as EmoDB and SAVEE. In fact, the reduction in RMSE is more prominent in EmoDB than in its balanced counterpart, the IITKGP dataset.

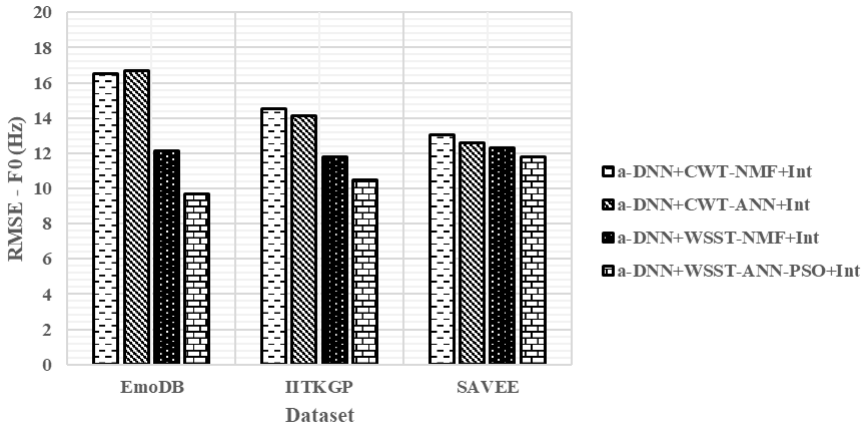
The effectiveness of the proposed algorithm for F0 mapping was already demonstrated by the reduced RMSE values for all the datasets considered for mapping. Further, for visualization purposes, a regression analysis was conducted to compare the ANN mapping methods after WSST-F0 and CWT-F0 decomposition, respectively. As a representative sample, the F0 regression plots for EmoDB are provided in Fig. 4.

Similar results were obtained for the regression performance in all datasets, keeping the same mapping schemes for spectral and intensity features. The difference in performance is most evident for fear in EmoDB, as seen in Fig. 4(c)-(d).

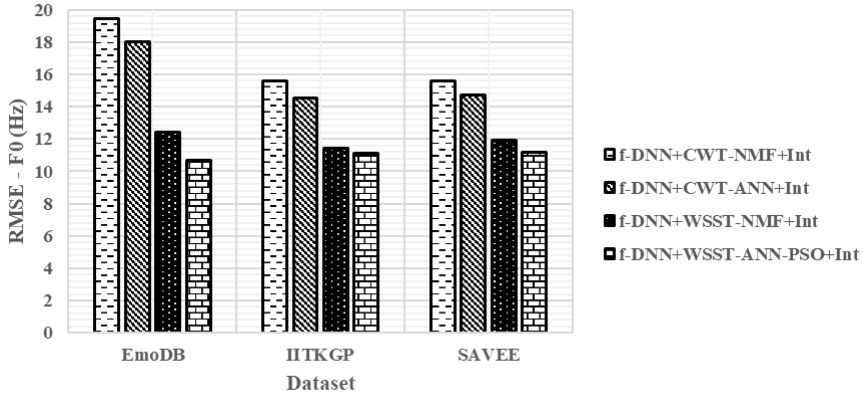
Whereas CWT-ANN mapping gives an overall R value of 0.698, WSST-ANN-PSO gives an R value of 0.925 for the same experimental conditions. Furthermore, the method performs equally well for happiness, as is evident in Fig. 4(f). In fact, the R value for happiness is even better than that for other emotions (overall $R = 0.945$).

Fig. 4 shows that the data are more focused and sharply coinciding with the regression line in the proposed method and are more scattered and spread out of the line in the CWT mapping methods. Additionally, it is observed that fear was the most difficult emotion to map using CWT-ANN-based training.

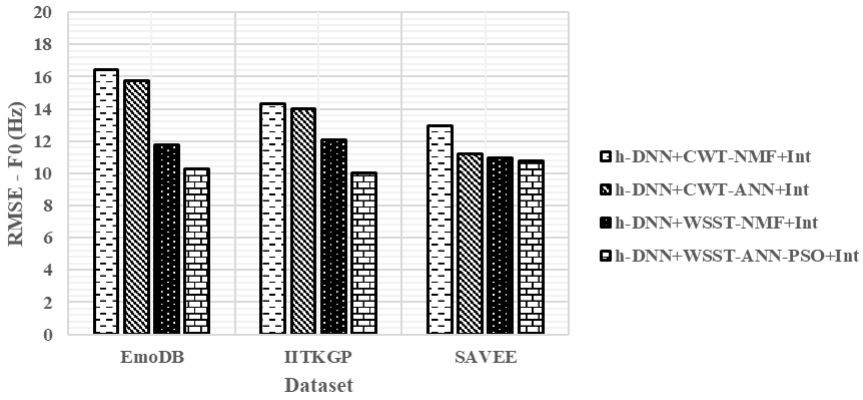
To evaluate the perceptive quality objectively, the signal PESQ scores were computed post conversion. The results are provided in Fig. 5. It is clear that a higher PESQ value is obtained for WSST-ANN-PSO-based conversion in all datasets. A higher PESQ value is obtained for anger and fear, with a maximum perception quality of 3.8 for fear in IITKGP. Almost consistent PESQ scores are obtained for happiness by both WSST-based methods. The better perception quality can be attributed to the reduced energy smearing in WSST transformation combined with optimized parameter mapping. Across datasets, PESQ values for the proposed method are almost consistent for anger. The consistency across dataset PESQ scores is reflected in all emo-



(a) Neutral to Anger



(b) Neutral to Fear



(c) Neutral to Happiness

FIGURE 3. **Comparison of RMSE- F0 (Hz) for conversion from neutral utterance to anger, fear and happiness. The baselines are represented as *DNN+CWT-NMF+Int*, *DNN+CWT-ANN+Int*, *DNN+WSST-NMF+Int*. a, f and h prefixes denote anger, fear and happiness respectively.**

tions, except fear. For fear, higher scores are obtained in IITKGP.

The MCD, F0-RMSE, and PESQ values as obtained above highlight the effectiveness of the *DNN + WSST-ANN-PSO + Int* framework for emotion conversion in all datasets. In the multiple conversion experiments, it was found that spectral mapping captures the variations in the low frequency region more effectively for all emotions. In addition, the variations

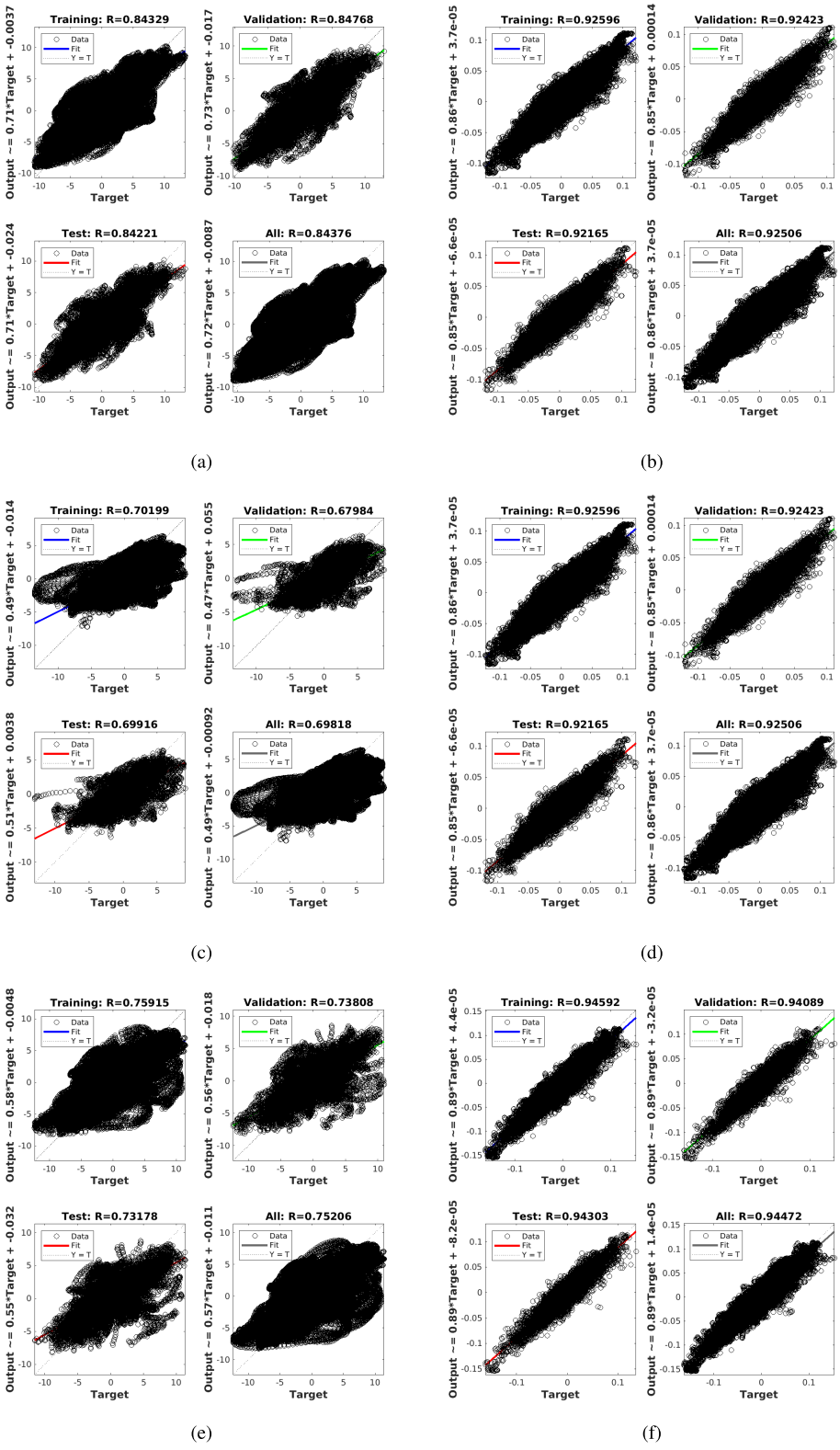
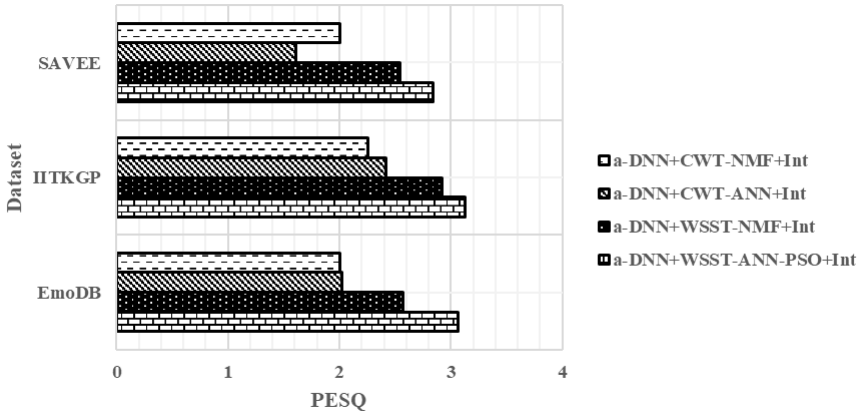
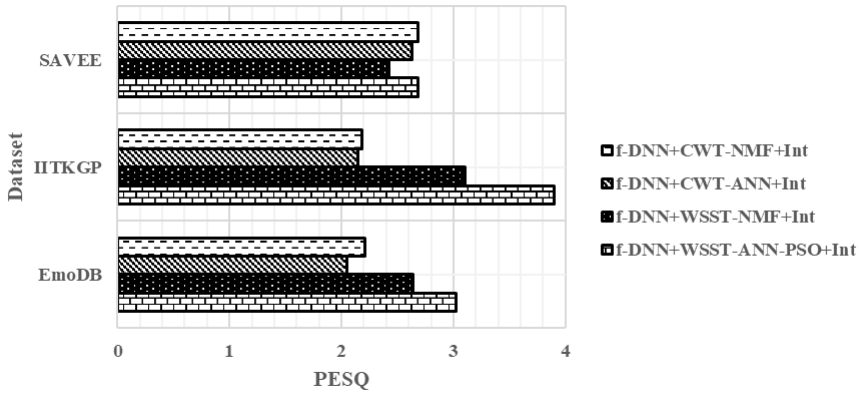


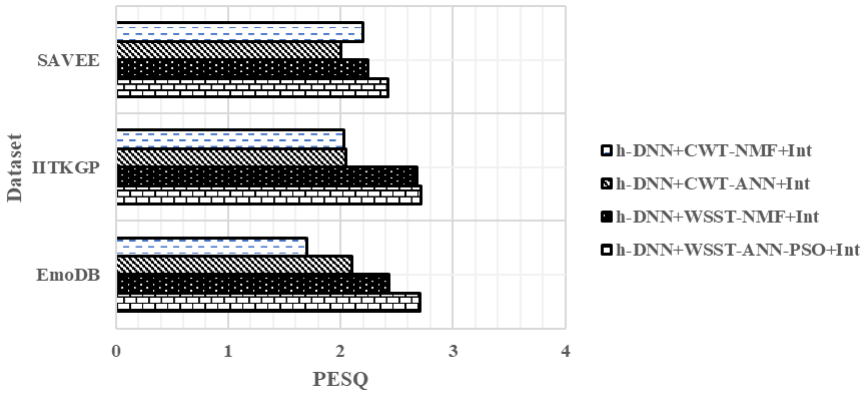
FIGURE 4. Regression performance with respect to F_0 mapping: (a)-(b) depict F_0 regression plot for anger, (c)-(d) that for fear, (e)-(f) for happiness in *DNN+CWT-ANN+Int* without parameter optimization of neural network and *DNN+WSST-ANN-PSO+Int* based mapping respectively. For illustration purpose, training/test data from EmoDB has been used for the plots.



(a) PESQ Scores - Anger



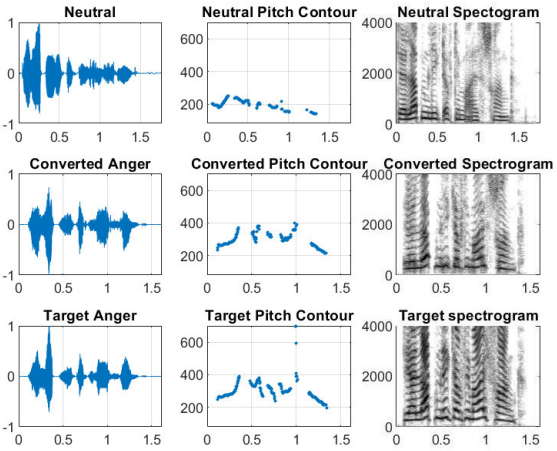
(b) PESQ Scores - Fear



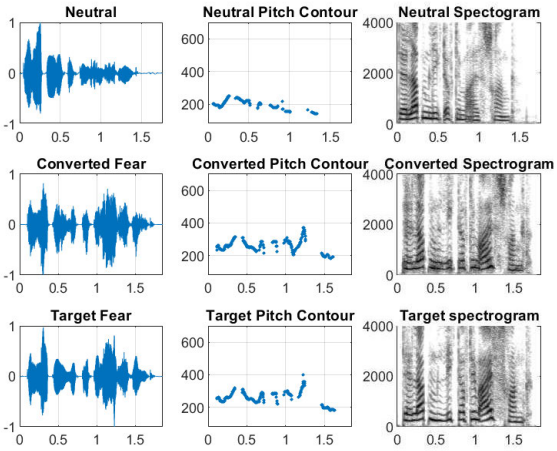
(c) PESQ Scores - Happiness

FIGURE 5. **Comparison of PESQ scores between proposed framework and state-of-the-art baselines: (a) gives the scores obtained for anger, (b) for fear and (c) for happiness across all datasets. The baselines are represented as *DNN+CWT-NMF+Int*, *DNN+CWT-ANN+Int*, *DNN+WSST-NMF+Int*. a, f and h prefixes denote anger, fear and happiness respectively.**

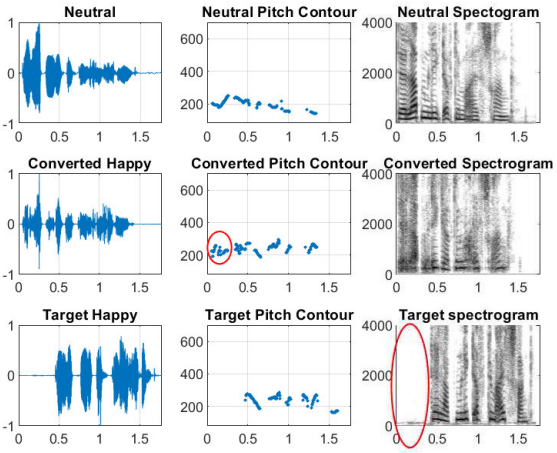
in higher formant regions for anger and happiness are well portrayed in IITKGP. Further, for illustration purposes and to obtain a visual perspective of the pitch and spectral mapping achieved by the proposed scheme, Fig. 6 is plotted with a single utterance from each of the datasets. In this figure, the pitch contour more closely follows the target for all emotions in IITKGP. In EmoDB, the duration of the utterances is longer and there are more breaks between them. The initial inflexions in happiness where there is an onset of a breathy voice quality (marked in red) are not clearly captured by the proposed method. This is interesting, because the F0 regression plot showed a better mapping performance in the case of happiness. Because F0 mapping is performed by taking instantaneous values and subsequently conducting interpola-



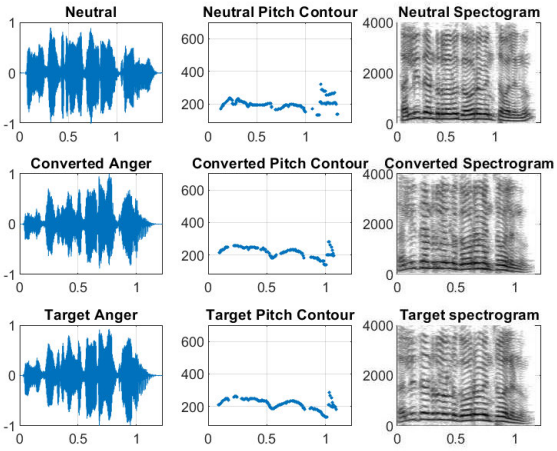
(a) Neutral to Anger-EmoDB



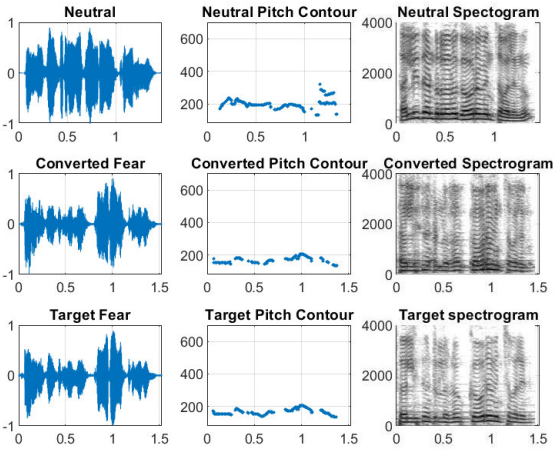
(b) Neutral to Fear-EmoDB



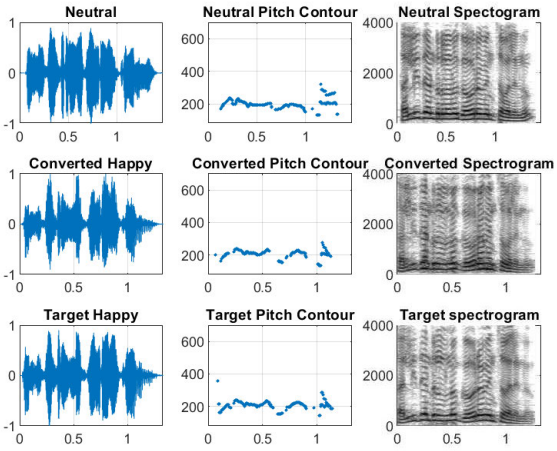
(c) Neutral to Happiness-EmoDB



(d) Neutral to Anger-IITKGP

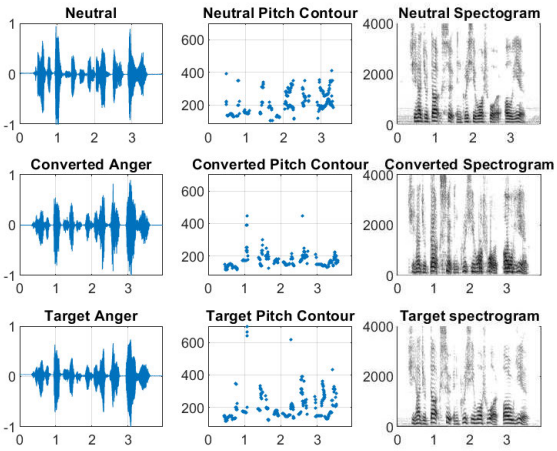


(e) Neutral to Fear-IITKGP

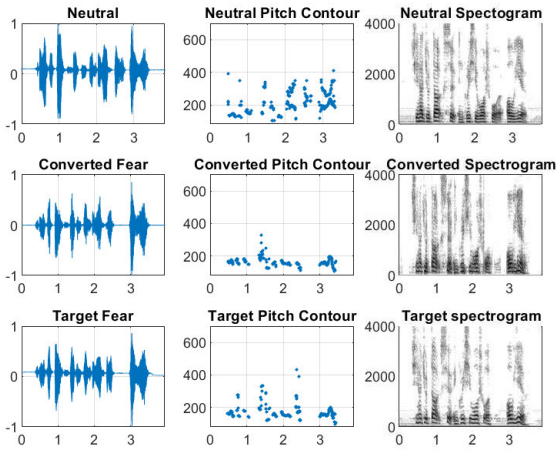


(f) Neutral to Happiness-IITKGP

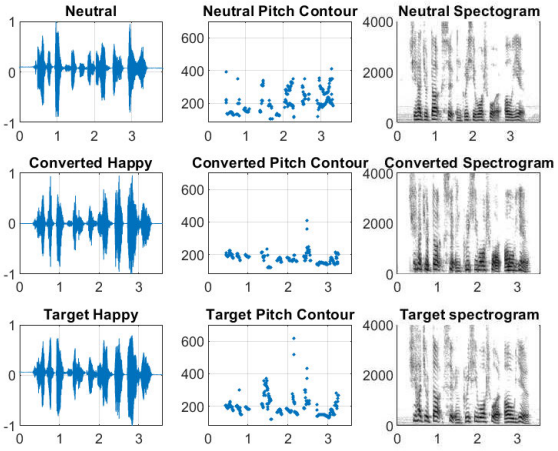
FIGURE 6. Pitch contours and spectrograms of neutral, converted and target speech utterances. Fig. (a)-(c) represent anger, fear and happiness conversion in EmoDB for German utterance 'Der Lappen liegt auf dem Eisschrank', (d)-(f) represent that for IITKGP for the Telugu utterance 'Talli tandrilenu goruavinchi vellanu' and (g)-(i) represent that for SAVEE for the utterance 'She had your dark suit in greasy wash water all year' in English respectively.



(g) Neutral to Anger-SAVEE



(h) Neutral to Fear-SAVEE



(i) Neutral to Happiness-SAVEE

FIGURE 6. ***(Continued.)*** **Pitch contours and spectrograms of neutral, converted and target speech utterances. Fig. (a)-(c) represent anger, fear and happiness conversion in EmoDB for German utterance 'Der Lappen liegt auf dem Eisschrank', (d)-(f) represent that for IITKGP for the Telugu utterance 'Talli tandrilenu goruavinchi vellanu' and (g)-(i) represent that for SAVEE for the utterance 'She had your dark suit in greasy wash water all year' in English respectively.**

tion, the silences and breathy voice quality are not considered for modification. In SAVEE, the variations in pitch contour are more effectively captured for fear. This is also evident in Fig. 3(b) where the RMSE values are comparatively lower for fear than for the other two emotions. Similar trends were obtained for other utterances also across datasets.

It is worthwhile to note that the happiness contour is most effectively mapped in IITKGP, although the F0-RMSE values are comparable for all three datasets, as shown in Fig. 3. Observation of the overall manner in which happiness is expressed in training samples from the IITKGP dataset shows that the breathy voice quality is less prominent and a ''pleasant'' feel is provided to the Telugu utterances. The method can be further refined by adding voice quality parameters, such as jitter and shimmer, to the mapping strategy.

B. SUBJECTIVE EVALUATION AND ANALYSIS

The strength of the *DNN + WSST-ANN-PSO + Int* framework for emotion conversion was evaluated through objective comparison metrics, as discussed above. Additionally, testing of the framework in terms of the perceptive quality of emotional expression was required. Ten listeners in the 20–30 year-old age group participated in perception tests using utterances selected randomly from each dataset. A total of 72 utterances were selected for testing (4 methods × 3 emotions × 3 datasets) × 2 (male/female). Because the German language was not familiar to listeners in the experiment site, for EmoDB alone the evaluation was conducted in two different phases, that is, with in-house non-native, as well as native German listeners. The factors considered in the evaluation were the CMOS and speaker similarity. At the beginning of each test, the utterance corresponding to target emotion is played, followed by utterances with synthesized emotion, in random order. The listeners were asked to score

TABLE 6. Subjective comparison between individual and combined parameter mapping across datasets.

Method		DNN+ Int	WSST-ANN-PSO+ Int	DNN+ WSST-ANN-PSO+Int	DNN+ Int	WSST-ANN-PSO+ Int	DNN+ WSST-ANN-PSO+Int
Mapped Features		MGCEP+Int	F_0+Int	MGCEP+F_0+Int	MGCEP+Int	F_0+Int	MGCEP+F_0+Int
Dataset	Emotion	Average CMOS			Average Speaker Similarity		
EmoDB (non-native)	Anger	2.6	2.7	3.5	2.6	2.3	3.8
	Fear	2.9	2.8	3.5	3.0	2.6	3.9
	Happiness	3.2	3.2	3.3	3.3	3.0	3.9
EmoDB (native)	Anger	2.2	2.2	3.4	2.2	2.2	3.4
	Fear	3.0	3.4	4.5	2.8	3.4	3.8
	Happiness	2.6	3	**2.8**	3.0	3.2	**2.8**
IITKGP	Anger	3.0	2.9	3.7	3.1	2.8	4.6
	Fear	2.7	2.8	3.6	2.5	2.4	3.9
	Happiness	2.8	2.8	3.6	3.4	2.6	3.9
SAVEE	Anger	2.7	2.8	3.5	2.8	2.7	4.0
	Fear	2.7	2.5	3.5	2.7	2.3	3.5
	Happiness	2.9	2.9	3.3	3.1	2.4	3.5

the synthesized utterance simultaneously in terms of two measures, similarity to target (CMOS evaluation) and speaker (speaker similarity scoring).

Because the proposed framework involves mapping of spectral, F0, and intensity features for emotion conversion, the individual contribution of spectral and F0 features in terms of objective evaluation criteria was already discussed. Since emotion is also subject to the listener's perception, the same emotions needed to be studied separately in terms of subjective evaluation. The first step in perception testing involved an investigation to determine the contribution of these features in expression mapping individually. For this purpose, a listening test was conducted using the same strategy as discussed above. Eighteen utterances from male and female speakers were considered, making a total of 36 utterances for testing (2 methods $\times$ 3 emotions $\times$ 3 datasets) $\times$ 2 (male/female). A comparison was made of the utterances synthesized by means of individual methods, viz. *DNN + Int*, *WSST-ANN-PSO + Int* and the proposed *DNN + WSST-ANN-PSO + Int* for mapping of features viz. MGCEP + Int, F_0 + Int and the combined *MGCEP* + F_0 + *Int* respectively. The results are recorded in Table. 6.

In Table. 6, rather than single parameter mapping, the modification of both spectral and F0 parameters and subsequent resynthesis yield a better CMOS and speaker similarity. For German utterances, native speakers gave slightly lower scores for happiness in terms of both CMOS and speaker similarity. Further, although almost identical CMOS scores are obtained for individual mapping of MGCEP or F0, the speaker similarity scores are slightly better for MGCEP mapping. This enhancement in similarity can be attributed to the utilization of speaker adaptation seed data for DNN-based feature mapping. The speaker identity is learned from the adaptation data provided for spectral mapping.

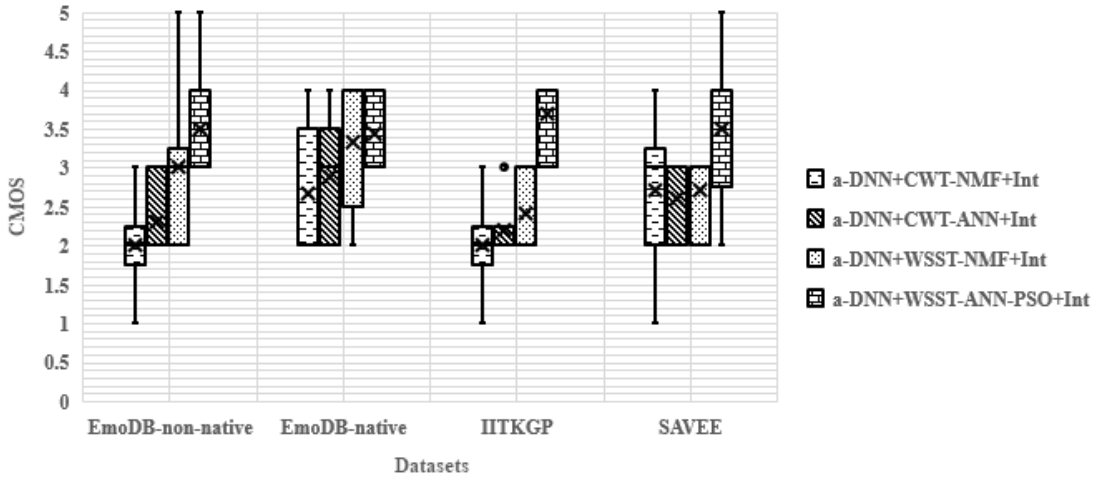
Since it was already established that the combined mapping of spectral, F0, and intensity parameters yields a better perceptive quality, an overall CMOS and speaker similarity scoring and comparison with various methods were conducted; the results are illustrated in Figs. 7 and 8.

The CMOS scores show that the proposed model surpassed the other methods in terms of similarity with the target for all emotions considered across all datasets. A comparison among different emotions shows that the maximum agreement in perceptive quality was obtained for fear in EmoDB. Comparable scores for anger were obtained for both EmoDB and IITKGP. However, for happiness, the scores for the Telugu dataset were slightly better. This may be because the voice quality for happiness was captured better in Telugu.

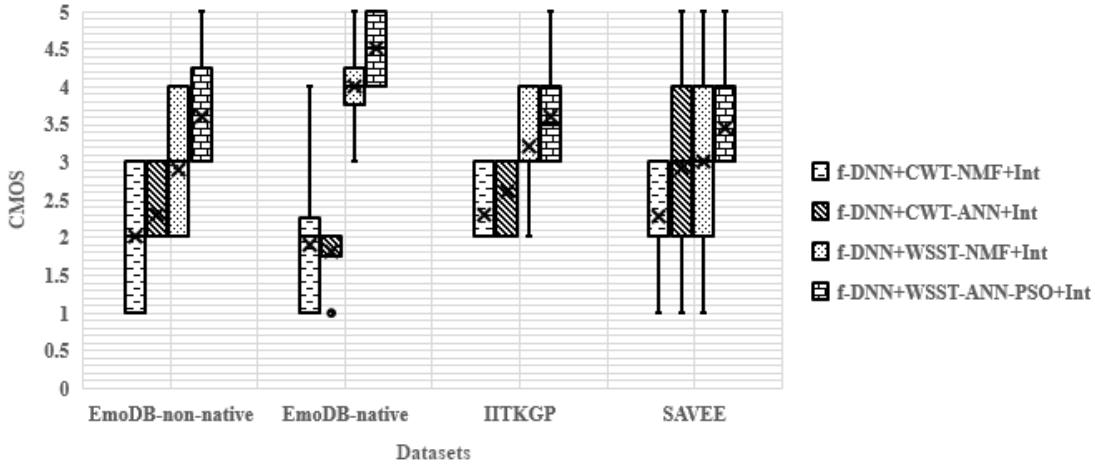
In Fig. 7(a), the mean value (marked as "x") is higher for the WSST-based methods. Additionally, the distribution tends more toward the upper quartile in the proposed framework across datasets. In the case of the German language samples, native listeners gave better scores for fear than for the other emotions. For happiness, a lower CMOS was recorded by native listeners. Overall, the listeners noted a closer similarity between the converted and target utterances for both WSST-based mapping methods than for that utilizing CWT. Thus, using the proposed mapping framework, an average CMOS value of 3.58 (anger), 3.78 (fear), and 3.35 (happiness) were obtained across datasets. Because the data are skewed, whiskers are not visible in cases where the lower quartile coincides with the minimum value or where the upper quartile falls into the maximum.

In Fig. 7(b), a similar trend can be observed for fear. In particular, considering the IITKGP dataset, although the maximum values are the same for both WSST-based methods, the distribution tends more toward the upper quartile in WSST-ANN-PSO mapping; the opposite tendency is seen for WSST-NMF. In Fig. 7(c), slightly higher CMOS scores are obtained for happiness in IITKGP, but the inter-quartile range is smaller in EmoDB for the proposed technique. Although SAVEE shows almost consistent scores for the proposed method across all emotions, the distribution moves totally toward the upper quartile in fear, which shows agreement with the higher CMOS for fear.

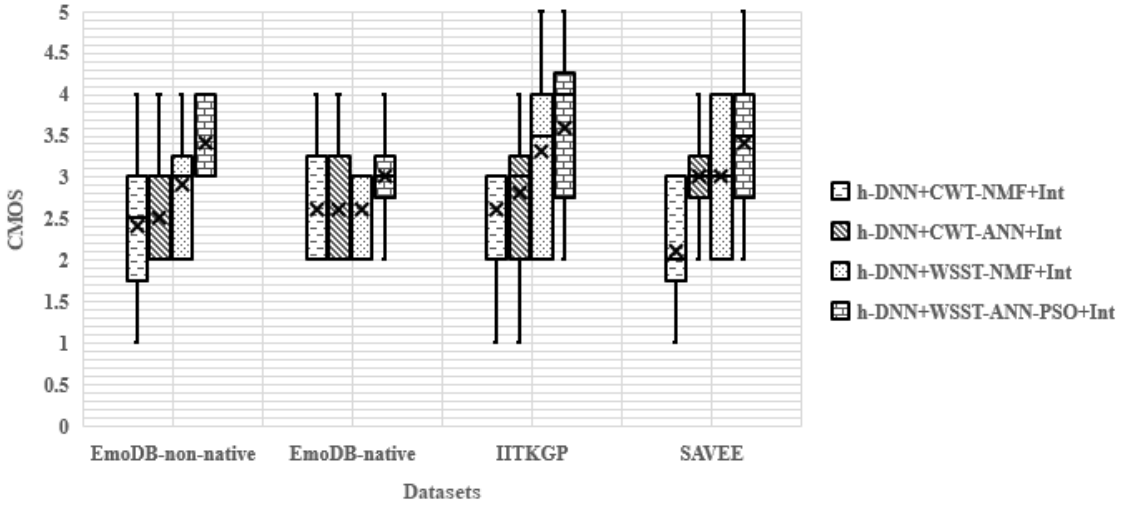
Higher speaker similarity scores were also obtained using the proposed model, as is evident in Fig. 8. A comparison of the datasets used for experimentation shows that



(a) CMOS Comparison - Anger

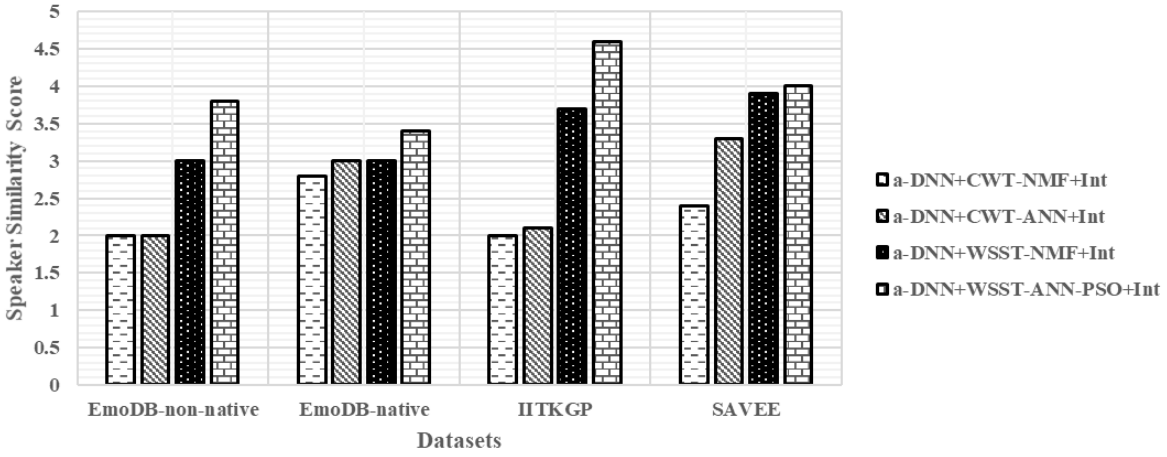


(b) CMOS Comparison - Fear

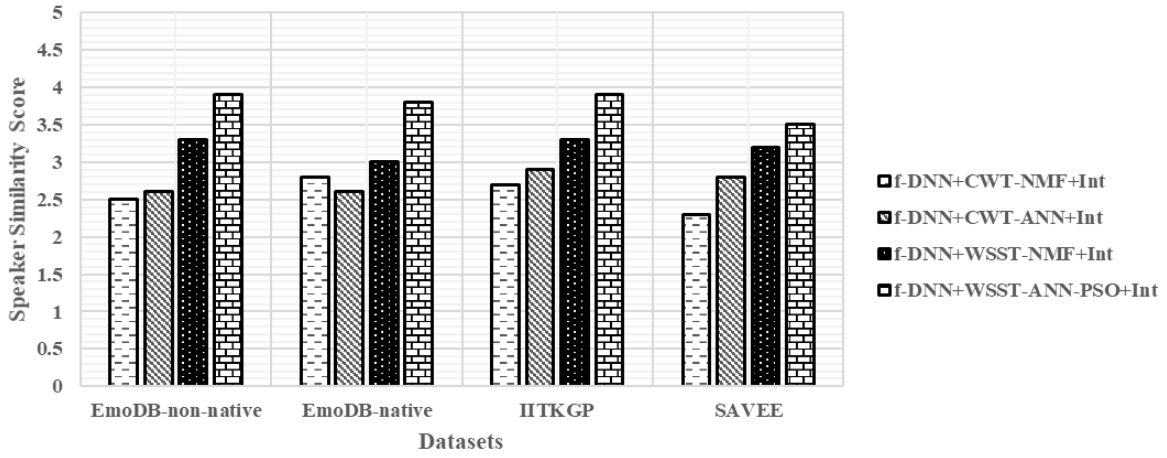


(c) CMOS Comparison - Happiness

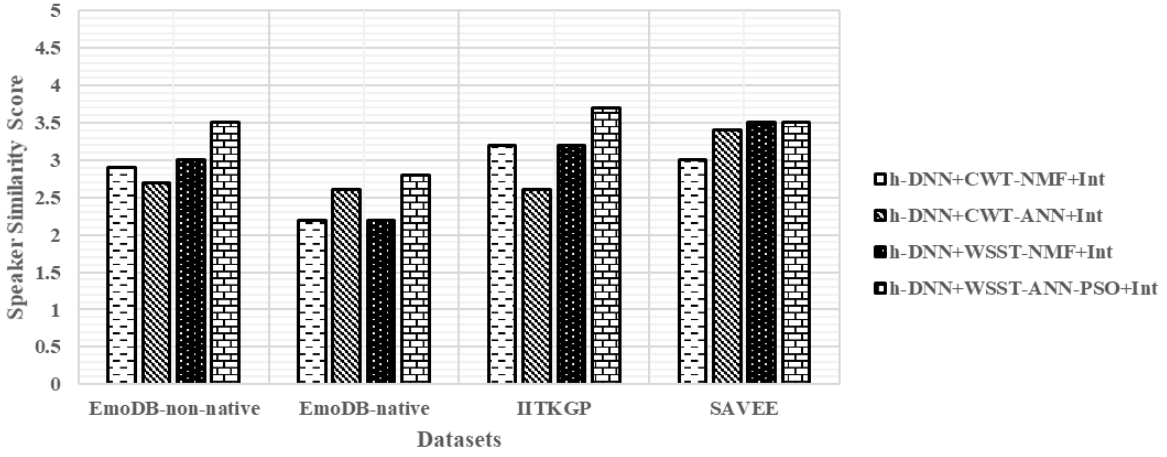
FIGURE 7. **Comparison of CMOS scores between proposed framework and state-of-the-art baselines: (a) gives the scores obtained for anger, (b) for fear and (c) for happiness across all datasets. The baselines are represented as *DNN+CWT-NMF+Int*, *DNN+CWT-ANN+Int*, *DNN+WSST-NMF+Int*. a, f and h prefixes denote anger, fear and happiness respectively.**



(a) Speaker Similarity Scores - Anger



(b) Speaker Similarity Scores - Fear



(c) Speaker Similarity Scores - Happiness

FIGURE 8. Comparison of speaker similarity scores between proposed framework and state-of-the-art baselines: (a) gives the scores obtained for anger, (b) for fear and (c) for happiness across all datasets. The baselines are represented as *DNN+CWT-NMF+Int*, *DNN+CWT-ANN+Int*, *DNN+WSST-NMF+Int*. a, f and h prefixes denote anger, fear and happiness respectively.

the maximum speaker similarity for the proposed method is obtained in IITKGP (anger) and EmoDB and IITKGP yield identical similarity scores for the other emotions. Additionally, native listeners gave slightly lower speaker similarity scores for the proposed model for both anger and happiness, although they were still higher than the corresponding values for the other methods. The scores are slightly lower in SAVEE, as the adaptation data are slightly fewer owing to the reduced size of the dataset. However, an average similarity score above 3 is obtained for all the emotions using the proposed framework. Considering the values for each emotion across datasets, average speaker similarity scores of 3.95 (anger), 3.78 (fear), and 3.38 (happiness) were obtained across datasets. Additionally, an overall average CMOS value of 3.57, and a speaker similarity score of 3.70 was obtained across datasets.

The listeners recorded slightly lower scores for SAVEE and expressed concern that speaker identity is comparatively difficult to establish, because all the speakers are males with approximately the same type of voice and accent. Further, the degree of acceptance in representing happiness by multiple speakers and datasets was average.

VIII. CONCLUSION AND FUTURE SCOPE

A hybrid network model for speaker-adaptive emotion conversion with combined spectral, F0, and intensity mapping was proposed in this work. The proposed method, in a comparative evaluation with state-of-the-art methods, yielded an enhanced performance according to all objective and subjective evaluation criteria. It was also found that, rather than individual parameter mapping, combined mapping of F0 and spectrum yielded better results in all the datasets considered in this work. Throughout the experimentation, effort was invested in preserving speaker identity as much as possible while deriving an acceptable transformation for anger, fear, and happiness. The combined *DNN + WSST-ANN-PSO + Int* mapping scheme can be extended to other languages, where speaker resources for expression training are limited.

The requirement of parallel neutral-target data was one of the challenges faced in this work. The method can be extended to include non-parallel source-target training data which can cater even for low-resource languages. Additionally, because simulated emotions are used for feature training, recognition of the appropriate emotion from the dataset itself is challenging. In such cases, where listening tests are highly subjective and tedious to conduct, the development of an automated evaluation system for CMOS testing could lead to better reliability in scores with minimal manual labor. In the future, by utilizing the learning gained from the multilingual experiments, the framework can be extended to conversational dialogue systems in multiple languages.

ACKNOWLEDGMENT

The authors sincerely thank all the listeners for their participation in the perception tests. Further, the authors express gratitude to Rahul Kumar Yadav, IHP Technologiepark, Frankfurt, Germany, for his efforts in co-ordinating the perception tests for German listeners.

REFERENCES

[1] A. H. P. Sarkar, A. K. Dutta, M. G. Reddy, D. M. Harikrishna, P. Dhara, R. Verma, N. P. Narendra, S. B. S. Kr, J. Yadav, and K. S. Rao, "Designing prosody rule-set for converting neutral TTS speech to storytelling style speech for indian languages: Bengali, Hindi and Telugu," in *Proc. IC*, 2014, pp. 473–477.

[2] R. Verma, P. Sarkar, and K. S. Rao, "Conversion of neutral speech to storytelling style speech," in *Proc. 8th Int. Conf. Adv. Pattern Recognit. (ICAPR)*, Jan. 2015, pp. 1–6.

[3] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1137–1144, Jul. 2006.

[4] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," *Int. J. Speech Technol.*, vol. 22, no. 3, pp. 497–510, Sep. 2019, doi: 10.1007/s10772-018-09572-8.

[5] S. Lalitha and S. Tripathi, "Emotion detection using perceptual based speech features," in *Proc. IEEE Annu. India Conf. (INDICON)*, Dec. 2016, pp. 1–5.

[6] P. M. Krishna, R. P. Reddy, V. Narayanan, S. Lalitha, and D. Gupta, "Affective state recognition using audio cues," *J. Intell. Fuzzy Syst.*, vol. 36, no. 3, pp. 2147–2154, Mar. 2019.

[7] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2014, pp. 1–10.

[8] Anon. (2013). *Technology Development for Indian Languages Programme*. [Online]. Available: http://tdil.mit.gov.in/AboutUs.aspx

[9] Z. Wu and H. Li, "Voice conversion versus speaker verification: An overview," *APSIPA Trans. Signal Inf. Process.*, vol. 3, p. e17, Dec. 2014.

[10] J. Cahn, "The generation of affect in synthesized speech," *J. Amer. Voice I/O Soc.*, vol. 8, pp. 1–19, Jul. 1990.

[11] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, May 1996, pp. 373–376.

[12] J. P. Cabral and L. C. Oliveira, "Emovoice: A system to generate emotions in speech," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 1798–1801.

[13] D. Govind and S. R. M. Prasanna, "Dynamic prosody modification using zero frequency filtered signal," *Int. J. Speech Technol.*, vol. 16, no. 1, pp. 41–54, Mar. 2013.

[14] S. Vekkot and S. Tripathi, "Significance of glottal closure instants detection algorithms in vocal emotion conversion," in *Proc. Int. Workshop Soft Comp. Appl. (SOFA)*. Cham, Switzerland: Springer, 2016, pp. 462–473.

[15] H. K. Vydana, S. R. Kadiri, and A. K. Vuppala, "Voweccation for emotion conversion," *Circuits, Syst., Signal Process.*, vol. 35, no. 5, pp. 1643–1663, May 2016.

[16] S. Vekkot and S. Tripathi, "Inter-emotion conversion using dynamic time warping and prosody imposition," in *Proc. ISTA*, 2016, pp. 913–924.

[17] A. Haque and K. S. Rao, "Modification of energy spectra, epoch parameters and prosody for emotion conversion in speech," *Int. J. Speech Technol.*, vol. 20, no. 1, pp. 15–25, Mar. 2017.

[18] S. Vekkot, "Building a generalized model for multi-lingual vocal emotion conversion," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 576–580.

[19] S. Vekkot and S. Tripathi, "Vocal emotion conversion using WSOLA and linear prediction," in *Proc. 19th Int. Conf. Speech Comput. (SPECOM)*, 2017, pp. 777–787.

[20] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003, pp. 2401–2404.

[21] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[22] Z. Inanoglu and S. Young, "A system for transforming the emotion in speech: Combining data-driven conversion techniques for prosody and voice quality," in *Proc. ISCA*, 2007, pp. 490–493.

[23] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 6, pp. 1394–1405, Aug. 2010.
[24] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *Amer. J. Signal Process.*, vol. 2, no. 5, pp. 134–138, Dec. 2012.
[25] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Commun.*, vol. 54, no. 1, pp. 134–146, Jan. 2012.
[26] J. Přibil and A. Přibilová, "GMM-based evaluation of emotional style transformation in Czech and Slovak," *Cognit. Comput.*, vol. 6, no. 4, pp. 928–939, Dec. 2014.
[27] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[28] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *Proc. INTERSPEECH*, 2011, pp. 669–672.
[29] S. Vekkot and D. Gupta, "Emotion conversion in Telugu using constrained variance GMM and continuous wavelet transform-F_0," in *Proc. IEEE Region Conf. (TENCON)*, Oct. 2019, pp. 991–996.
[30] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 912–921, Jul. 2010.
[31] S. R. Krothapalli, J. Yadav, S. Sarkar, S. G. Koolagudi, and A. K. Vuppala, "Neural network based feature transformation for emotion independent speaker identification," *Int. J. Speech Technol.*, vol. 15, no. 3, pp. 335–349, Sep. 2012.
[32] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 1969–1973.
[33] J. Yadav and K. S. Rao, "Prosodic mapping using neural networks for emotion conversion in Hindi Language," *Circuits, Syst., Signal Process.*, vol. 35, no. 1, pp. 139–162, Jan. 2016.
[34] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–5.
[35] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17230–17246, 2019.
[36] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2013, pp. 104–108.
[37] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5175–5179.
[38] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," 2017, *arXiv:1711.05447*. [Online]. Available: http://arxiv.org/abs/1711.05447
[39] K. S. Rao and A. K. Vuppala, "Non-uniform time scale modification using instants of significant excitation and vowel onset points," *Speech Commun.*, vol. 55, no. 6, pp. 745–756, Jul. 2013.
[40] V. V. R. Vegesna, K. Gurugubelli, and A. K. Vuppala, "Prosody modification for speech recognition in emotionally mismatched conditions," *Int. J. Speech Technol.*, vol. 21, no. 3, pp. 521–532, Sep. 2018.
[41] S. L. Priya and D. Govind, "Significance of epoch identification accuracy in prosody modification for effective emotion conversion," in *Proc. Int. Symp. Signal Process. Intell. Recognit. Syst.* Singapore: Springer, 2018, pp. 334–346.
[42] S. Vekkot and D. Gupta, "Prosodic transformation in vocal emotion conversion for multi-lingual scenarios: A pilot study," *Int. J. Speech Technol.*, vol. 22, no. 3, pp. 533–549, Sep. 2019.
[43] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, Oct. 2011.
[44] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for mandarin," in *Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*, vol. 4, 2007, pp. 410–414.
[45] M. S. Ribeiro and R. A. J. Clark, "A multi-level representation of F0 using the continuous wavelet transform and the discrete cosine transform," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4909–4913.
[46] M. Vainio, A. Suni, and D. Aalto, "Continuous wavelet transform for analysis of speech prosody," in *Proc. TRASP*, 2013, pp. 78–81.
[47] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *Proc. 8th ISCA Workshop Speech Synth.*, 2013, pp. 78–81.
[48] H. Ming, D. Huang, M. Dong, H. Li, L. Xie, and S. Zhang, "Fundamental frequency modeling using wavelets for emotional voice conversion," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 804–809.
[49] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP J. Audio, Speech, Music Process.*, vol. 2017, no. 1, 2017, Art. no. 18.
[50] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data," in *Proc. INTERSPEECH*, Aug. 2017, pp. 3399–3403.
[51] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero, "Emotion transplantation through adaptation in HMM-based speech synthesis," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 292–307, Nov. 2015.
[52] H. Yang, W. Zhang, and P. Zhi, "A DNN-based emotional speech synthesis by speaker adaptation," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 633–637.
[53] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion to emotional speech based on three-layered model in dimensional approach and parameterization of dynamic features in prosody," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–6.
[54] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," 2018, *arXiv:1811.01174*. [Online]. Available: http://arxiv.org/abs/1811.01174
[55] S. Vekkot, D. Gupta, M. Zakariah, and Y. A. Alotaibi, "Hybrid framework for speaker-independent emotion conversion using i-vector PLDA and neural network," *IEEE Access*, vol. 7, pp. 81883–81902, 2019.
[56] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of F0 for speech emotion conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6830–6834.
[57] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks," *APSIPA Trans. Signal Inf. Process.*, vol. 8, pp. 1–11, Mar. 2019.
[58] B. Yegnanarayana, *Artificial Neural Networks*. New Delhi, India: PHI Learning, 2009.
[59] I. A. Basheer and M. Hajmeer, "Artificial neural networks: Fundamentals, computing, design, and application," *J. Microbiol. Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000.
[60] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[61] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8609–8613.
[62] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[63] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[64] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 243–261, Mar. 2011.
[65] D. Govind, S. L. Priya, S. Akarsh, B. G. Gowri, and K. P. Soman, "Improved epoch extraction from speech signals using wavelet synchrosqueezed transform," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2019, pp. 1–5.
[66] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61. Philadelphia, PA, USA: SIAM, 1992.
[67] I. Daubechies and S. Maes, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," in *Wavelets in Medicine and Biology*, A. Aldroubi and M. Unser, Eds. Boca Raton, FL, USA: CRC Press, 1996, pp. 527–546.
[68] Y. Zhao, H. Cui, H. Huo, and Y. Nie, "Application of synchrosqueezed wavelet transforms for extraction of the oscillatory parameters of subsynchronous oscillation in power systems," *Energies*, vol. 11, no. 6, p. 1525, 2018.

[69] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw. (ICNN)*, Perth, WA, Australia, vol. 4, Nov. 1995, pp. 1942–1948.
[70] D. Niu and M. Xing, "Research on neural networks based on culture particle swarm optimization and its application in power load forecasting," in *Proc. 3rd Int. Conf. Natural Comput. (ICNC)*, vol. 1, 2007, pp. 270–274.
[71] Z. A. Bashir and M. E. El-Hawary, "Applying wavelets to short-term load forecasting using PSO-based neural networks," *IEEE Trans. Power Syst.*, vol. 24, no. 1, pp. 20–27, Feb. 2009.
[72] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proc. 6th Int. Symp. Micro Mach. Hum. Sci.*, 1995, pp. 39–43.
[73] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization," *Swarm Intell.*, vol. 1, no. 1, pp. 33–57, Jun. 2007.
[74] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of Machine Learning*. 2010, pp. 760–766.
[75] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlemeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2011, pp. 1517–1520.
[76] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "IITKGP-SESC: Speech database for emotion analysis," in *Proc. IC*. Berlin, Germany: Springer, 2009, pp. 485–492.
[77] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proc. Int. Conf. Audio Visual Speech Process.*, 2009, pp. 53–58.
[78] E. Azarov, M. Vashkevich, and A. Petrovsky, "Instantaneous pitch estimation based on RAPT framework," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2012, pp. 2787–2791.
[79] *Perceptual Evaluation of Speech Quality PESQ: An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, document Rec. ITU-T P. 862, 2001.
[80] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.
[81] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (PESQ) the new ITU standard for end-to-end speech quality assessment Part I—Time-delay compensation," *J. Audio Eng. Soc.*, vol. 50, no. 10, pp. 755–764, 2002.
[82] V. Sethu, E. Ambikairajah, and J. Epps, "Empirical mode decomposition based weighted frequency feature for speech-based emotion classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 5017–5020.
[83] A. R. Avila, M. J. Alam, D. O'Shaughnessy, and T. Falk, "Investigating speech enhancement and perceptual quality for speech emotion recognition," in *Proc. INTERSPEECH*, Sep. 2018, pp. 3663–3667.
[84] D. Govind, S. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *Proc. INTERSPEECH*, 2011, pp. 2969–2972.
[85] D. Govind and S. R. M. Prasanna, "Expressive speech synthesis: A review," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 237–260, Jun. 2013.
[86] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Commun.*, vol. 102, pp. 54–67, Sep. 2018.

SUSMITHA VEKKOT (Member, IEEE) was born in 1982. She received the M.S. degree in digital communications networks from London Metropolitan University, U.K., in 2009. She is currently pursuing the Research degree with the Amrita School of Engineering, Bengaluru, India. She is also a Full-Time Ph.D. Scholar under Government of India's Visveswaraya Ph.D. Scheme. She has a teaching experience of over nine years. She has conducted and participated in number of short-term courses, seminars, and conferences conducted at the national/international level. She has published articles in several International conferences and journals of repute. Her research work is published in reputed international journals like IEEE Access, *International Journal of Speech Technology*, and so on. Her research interests include speech processing, applications of soft computing techniques in speech, voice conversion, emotion analysis and synthesis, and human–machine interaction. She is member of the Professional Technical Society International Speech Communication Association (ISCA).

DEEPA GUPTA was born in 1977. She received the Ph.D. degree in natural language processing (example-based machine translation) from the Department of Mathematics and Computer Application, IIT Delhi, in 2005. She worked as a Postdoctoral Researcher with FBKIRST (Center for Scientific and Technological Research), Trento, Italy. She is currently an Associate Professor with the Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru, India. Her research work is published in journals like *Information Processing and Management*, *Expert Systems With Applications*, IEEE Access, *International Journal of Speech Technology*, and so on. She has been guiding Ph.D. and graduate students, since 2009. She has given invited talks on machine learning and natural language processing in Government funded workshops. She has completed two government funded projects related to text plagiarism detection and speech recognition system for Kannada language in last five years. She is also involved in consultancy projects with industry. Her research interests are in sentiment analysis, clinical data mining, speech processing, and other areas in natural language processing.

MOHAMMED ZAKARIAH received the B.Sc. degree in computer science and engineering from Visvesvaraya Technological University, India, in 2005, and the master's degree in computer engineering from Jawaharlal Nehru Technological University, India, in 2007. He is currently a Researcher with the Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He has published more than 20 articles in various reputed journals. His research interests include bioinformatics, digital audio forensics, speech processing, cloud computing, multimedia, healthcare, and social media.

YOUSEF AJAMI ALOTAIBI (Senior Member, IEEE) received the B.Sc. degree in computer engineering from King Saud University, Riyadh, Saudi Arabia, in 1988, and the M.Sc. and Ph.D. degrees in computer engineering from the Florida Institute of Technology, FL, USA, in 1994 and 1997, respectively. From 1988 to 1992 and 1998 to 1999, he was a Research Engineer with Al-ELM Research and Development Corporation, Riyadh. From 1999 to 2008, he was an Assistant Professor with the College of Computer and Information Sciences, King Saud University, where he was also an Associate Professor, from 2008 to 2012. Since 2012, he has been a Professor with King Saud University. His research interests are digital speech processing, specifically speech recognition, and arabic language and speech processing.