Geravanchizadeh *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:31
https://doi.org/10.1186/s13636-021-00216-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Feature compensation based on the normalization of vocal tract length for the improvement of emotion-affected speech recognition

Masoud Geravanchizadeh*, Elnaz Forouhandeh and Meysam Bashirpour

Abstract

The performance of speech recognition systems trained with neutral utterances degrades significantly when these systems are tested with emotional speech. Since everybody can speak emotionally in the real-world environment, it is necessary to take account of the emotional states of speech in the performance of the automatic speech recognition system. Limited works have been performed in the field of emotion-affected speech recognition and so far, most of the researches have focused on the classification of speech emotions. In this paper, the vocal tract length normalization method is employed to enhance the robustness of the emotion-affected speech recognition system. For this purpose, two structures of the speech recognition system based on hybrids of hidden Markov model with Gaussian mixture model and deep neural network are used. To achieve this goal, frequency warping is applied to the filterbank and/or discrete-cosine transform domain(s) in the feature extraction process of the automatic speech recognition system. The warping process is conducted in a way to normalize the emotional feature components and make them close to their corresponding neutral feature components. The performance of the proposed system is evaluated in neutrally trained/emotionally tested conditions for different speech features and emotional states (i.e., *Anger*, *Disgust*, *Fear*, *Happy*, and *Sad*). In this system, frequency warping is employed for different acoustical features. The constructed emotion-affected speech recognition system is based on the Kaldi automatic speech recognition with the Persian emotional speech database and the crowd-sourced emotional multi-modal actors dataset as the input corpora. The experimental simulations reveal that, in general, the warped emotional features result in better performance of the emotion-affected speech recognition system as compared with their unwarped counterparts. Also, it can be seen that the performance of the speech recognition using the deep neural network-hidden Markov model outperforms the system employing the hybrid with the Gaussian mixture model.

Keywords: Emotion-affected speech recognition, Vocal tract length normalization, Frequency warping, Acoustic feature normalization

* Correspondence: geravanchizadeh@tabrizu.ac.ir
Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz 51666-15813, Iran

1 Introduction

Speech is the natural medium of communication for humans. In recent years, improvements in speech technology have led to a considerable enhancement in human-computer interaction. The applications of such technology are numerous, including speech and speaker recognition, interactive voice response (IVR), dictation systems, and voice-based command and control for robots, etc.

Despite all the recent advances in speech processing systems, often these systems struggle with issues caused by speech variabilities. Such variabilities in speech can occur due to speaker-dependent characteristics (e.g., the shape of the vocal tract, age, gender, health, and emotional states), environmental noise, channel variability, speaking rate (e.g., changes in timing and realization of phonemes), speaking style (e.g., read speech vs. spontaneous speech), and accent variabilities (e.g., regional accents or non-native accents) [1].

Over the last decades, automatic speech recognition (ASR) systems have progressed significantly. The function of these systems is to recognize the sequence of words uttered by a speaker. Speech recognizers could be used in many applications for more convenient human-machine interaction, including mobile phones, smart home devices, intelligent vehicles, medical devices, and educational tools.

It is known that speech variabilities such as emotions could affect speech recognition performance considerably. Although most of the research in this area has been focused on the recognition of speech emotions, limited works have been performed in the area of emotion-affected speech recognition (EASR).

Generally, in real-life applications, there is an incompatibility between training and testing conditions. The current approaches for the reduction of the mismatch between the speech sets of neutral training and emotional testing of the EASR system can be categorized into three main classes.

In the first class of approaches, called model adaptation, a re-training of acoustic models is achieved. The adaptation techniques in this group include maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) [2]. Vlasenko et al. [3] employed MLLR and MAP and applied them to the German emotional database (EMO-DB) [4] to improve the recognition performance. In an attempt to use a fast adaptation method, Pan et al. [5] used the MLLR technique to construct emotion-dependent acoustic models (AMs) by employing a small portion of the Chinese emotional database. Here, first, the Gaussian mixture model (GMM)-based emotion recognition is performed to improve the performance of the speech recognition by selecting an appropriate emotion-match model. Another study was accomplished by Schuller et al. [6] in the framework of model adaptation. They employed adaptation methods in a hybrid ASR which is constructed by a combination of an artificial neural network (ANN) and a hidden Markov model (HMM) to form the ANN-HMM ASR structure. Compared to a static adaptation strategy, it is observed that maximum improvement in recognition is obtained by a dynamic adaptation method. To remedy the influence of emotion in recognition, Ijima et al. [7] have involved the paralinguistic information into the HMM process, which resulted in style estimation and adaptation using the multiple regression HMM (MRHM M) technique.

In the second group of methods, some knowledge clues are added to the language model (LM) of the ASR system to reduce the mismatch between the neutral training and emotional testing conditions. This strategy was taken by Athanaselis et al. [8] who explained how an emotion-oriented LM can be constructed from the existing British national corpus (BNC). In their method, an increased representation of emotional utterances is obtained by, first, identifying emotional words in BNC using an emotional lexicon. Then, sentences containing these words are recombined with BNC to construct a corpus with a raised proportion of emotional material. The corpus is then used to design emotionally enriched LM to improve recognition performance with emotional utterances.

The third class of approaches to compensate mismatch of acoustic characteristics between neutral utterances and emotionally affected speech materials involves those that study acoustic and prosodic features intending to provide robust features to overcome the performance degradation in the EASR systems. In [9], the performance of the HMM-based emotional speech recognizer is improved by using 12 cepstral coefficients, the logarithm of energy, their first- and second-order delta parameters, and additional features, such as the pitch frequency, its slope, and per-speaker-syllable-z-normalization (PSSZN). The study in [10] examines the changes in formant frequencies and pitch frequency due to emotional utterances. The HMM-based speech recognizer uses one log-energy coefficient and cepstral coefficients plus their delta and delta-delta parameters as its typical feature vector. In this work, an emotion recognition process is first conducted to find more appropriate parameters to be included in the feature vector. The results show that adding supplementary features such as pitch and formant frequencies to the feature vector is useful in improving emotional speech recognition. In another study, Sun et al. [11] proposed a new feature, called F-ratio scale frequency cepstral coefficients (FFCCs) that employed Fisher's F-ratio to analyze the importance of frequency bands for enhancing the mel-frequency cepstral coefficient (MFCC)/perceptual linear prediction (PLP) filterbank design in emotional condition. The simulation results show that employing the optimized features increases the recognition performance of EASR as compared to the conventional features of MFCC or PLP in the sense of sentence error rate (SER).

The performance of a speech recognition system degrades when there is a mismatch between the set of speakers used to train the system and that used to recognize it. This mismatch arises due to the anatomical differences of various speakers, as reflected in the vocal tract structures among different speakers. The result is that the system trained on specific speakers will perform poorly in the presence of other speakers. Vocal tract length normalization (VTLN) is one of the approaches to reduce the mismatch between training data and recognition data in an ASR system. Lee et al. [12] performed pioneering works in utilizing the VTLN technique for diminishing the performance reduction in an ASR system which is caused by variation of vocal tract length among different speakers. The procedure of speaker normalization is based on warping the frequency axis of mel-filterbank linearly in the process of extracting mel-frequency cepstrum features. To this aim, first, the warping factor is estimated efficiently in a model-based maximum likelihood framework. Then, the normalization process is conducted by scaling the frequency axis of the speech signal with the calculated warping factor. The recognition results show the effectiveness of this procedure for telephone-based connected digit databases. In another approach to implement the frequency warping for VTLN, Panchapagesan et al. [13] proposed a novel transformation matrix to perform warping in the discrete-cosine transform (DCT) calculation stage of the MFCC feature for speaker normalization in the ASR system. Compared with other linear transformation approaches, employing the proposed transformation matrix had a lower computational load without modifying the standard MFCC feature extraction procedure. For presenting the effectiveness of the new linear transformation method for VTLN, the DARPA resource management (RM1) database was used [14].

Conceptually, for a person speaking emotionally, the anatomical features of the speaker regarding the structure of his/her vocal tract are changed compared to those of a neutral speaking person. This fact implies that compensating the emotion-related variabilities on a speech by the technique of VTLN could increase the speech recognizer performance in emotional conditions. To improve the recognition rate of emotional speech, Sheikhan et al. [15] neutralized the MFCC features by applying the VTLN technique for the emotional states of *Anger* and *Happy*. The frequency warping of MFCCs is accomplished after finding the most emotion-affected frequency range. Finally, the neutralized MFCCs are employed in an HMM-based speech recognizer trained with neutral speech utterances. The simulation results demonstrate that applying the frequency warping to both modules of mel-filterbank and DCT yields better recognition performance as compared to the case in which the warping is applied only to the individual modules.

The previous studies have focused on applying VTLN as a normalization tool to MFCCs as the most popular acoustic feature in the speech recognition framework. The strategy taken in the present work is the same as in [15] in that VTLN is used to normalize the acoustical features extracted from an emotional utterance. However, our work differs from [15] in some aspects. Here, the robustness of different features including MFCCs is investigated in various emotional states with/without employing the cepstral mean normalization (CMN) in the EASR system. Next, a study was conducted to find an optimal frequency range in which warping is performed in the VTLN method. Also, the technique of VTLN is applied to other acoustical features than MFCCs to develop more robust features which can be used in improving the performance of the EASR system. Another aspect of the present work concerns the use of the deep neural network (DNN) in the structure of speech recognizer. Due to the high performance of DNNs in the acoustic modeling over the classical GMMs, the VTLN method is also employed with the state-of-the-art DNN-HMM speech recognizer.

The paper is organized as follows. In Section 2, the proposed EASR system and the technique of VTLN are presented, which describe the concept of warping or normalizing speech features in detail. Section 3 provides the experiments and recognition results for speech materials from two known databases with neutral and different emotional states. The simulation results presented in this section include examining the effect of applying CMN in the feature extraction process, investigating the influence of using different ranges of frequency warping, and evaluating the performance of various frequency warping methods for the GMM-HMM/DNN-HMM EASR system. The concluding remarks are given in Section 4.

2 Methods

2.1 Emotion-affected speech recognition system

The overall structure of the proposed EASR system is depicted in Fig. 1. Here, different emotional utterances serve as input to the system which is then converted into a sequence of acoustic features by the unit of feature extraction. However, the recognition rate of an ASR system trained with neutral speech degrades when features of emotional speech are fed into the system. This calls for a procedure for the normalization of acoustic

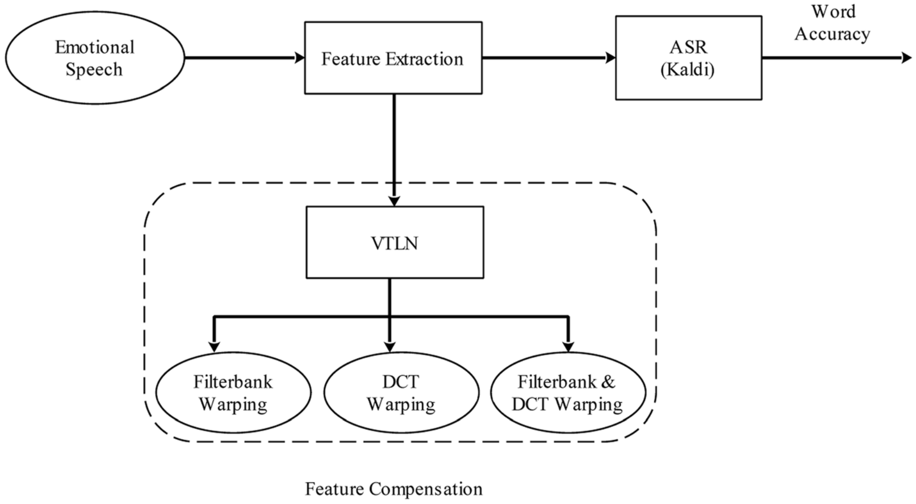


Fig. 1 The general structure of the proposed EASR system

features before they are used by the speech recognizer system. To this aim, the technique of VTLN is adopted in feature extraction to alleviate the effects of emotion in the speech recognition process. The VTLN approach can be performed either by frequency warping in filterbank, DCT unit, or both. After the process of feature normalization, the features are given to a speech recognizer as the back-end processing stage. In this work, the Kaldi speech recognizer [16] trained with neutral utterances of the Persian and English datasets [17, 18] is employed as the baseline ASR system.

2.2 Feature extraction

Feature extraction aims to find a set of feature vectors that are capable to capture the essential information as much as possible from the input speech signal. An ideal feature vector for emotional speech recognition application should maximize the discriminating ability of speech classes (e.g., phonemes) while it should not be affected by speaker-specific characteristics such as shape and length of the vocal tract.

Little research has been conducted on the robustness and suitability of different acoustic features in the emotional speech recognition framework. However, the studies made in the field of automatic speech recognition reveal that the most notable acoustic features are MFCC [19], modified mel-scale cepstral coefficient (M-MFCC) [20], exponential logarithmic scale (ExpoLog) [20], gammatone filterbank cepstral coefficient (GFCC) [21], linear prediction cepstral coefficient (LPCC) [22], relAtive specTrAl perceptual linear prediction (RASTA-PLP) [23], and power normalized cepstral coefficient (PNCC) [24].

It has been shown that M-MFCC and ExpoLog have performed better than MFCC in speech recognition under stress conditions [20]. The extraction procedure of these features is similar to MFCC, but it differs from that of MFCC in the frequency scaling of the filterbank.

GFCC was introduced as a robust feature for speech recognition in a noisy environment [21]. The process of GFCC feature extraction is based on the gammatone filterbank, which is derived from psychophysical observations of the auditory periphery.

PNCC is one of the acoustic features which provides notable results for the recognition of speech in noisy and reverberant environments [24]. The extraction of the PNCC feature is inspired by human auditory processing.

In this paper, MFCC, M-MFCC, ExpoLog, GFCC, and PNCC are employed as auditory features in the EASR system.

2.3 Feature normalization

As it was pointed out earlier, the mismatch between training and recognition phases causes performance degradation in ASR systems [25]. One of the sources of this mismatch can be associated with various speakers having vocal tracts with different anatomical features. Previous studies have shown that the acoustic and articulatory characteristics of speech are affected by the emotional content of the speech. There is evidence that, when a typical speaker speaks emotionally, the position of the tip of the tongue, jaw, and lips are changed, and this, in turn, modifies the acoustic features such as formant frequencies [26, 27]. This implies that the vocal tract length variation can be considered as a function of the emotional state of a person [28]. This, in turn, means that during the recognition of emotional speech, techniques are needed to decrease the influence of vocal tract length variations that arise from the emotional state of the speaker. Among different approaches that can be considered, the VTLN method is employed in this work as a way to remedy the mismatch problem in speech recognition applications.

2.3.1 Methods of VTLN

The VTLN technique views the main difference between two speakers as a change in the spectral content of acoustical features due to the differences in vocal tract length between speakers [29]. The idea of VTLN in speech recognition can be extended to the emotional speaking task, where the difference between emotional and neutral speech is associated with the variation of the frequency axis, originating from the vocal tract length differences of emotional and neutral speaking styles.

To cope with the mismatch problem between neutrally training and emotionally testing of an ASR system, the VTLN technique provides a warping function by which the frequency axis of the emotional speech spectrum is transformed to the frequency axis of the neutral speech spectrum. The normalization procedure can be performed by linear or nonlinear frequency warping functions such as piecewise linear, exponential functions, etc. [12]. These functions operate based on a warping parameter, which compresses or expands the speech spectra as follows [29]:

$$S_{\text{neutral}}(f) = S_{\text{emotional}}\left(f'(\alpha, f)\right) \qquad (1)$$

Here, f' is the frequency warping operation applied to the frequency axis of the emotional speech spectrum using α as the warping factor.

Most auditory-based acoustic features employ some sort of frequency decomposition (e.g., using filterbanks) and decorrelation of spectral features (e.g., using DCT processing) in their computations. This means that the warping of frequencies can be applied in the filterbank and/or DCT processing stage(s) of the acoustic feature extraction. Figure 2 represents the general block diagram of employing frequency warping in the filterbank and/or DCT domain(s) to compute the corresponding warped features of MFCC [19], M-MFCC [20], ExpoLog [20], GFCC [21], and PNCC [24] in one or both of the domains. For comparison purposes, the dashed boxes represent the optional cases of no-frequency warping which are used to generate the conventional features in their unwarped form. Intermediate operations performed for each feature extraction method are also illustrated. The warping strategies are discussed in detail below.

2.3.1.1 Frequency warping in the filterbank domain

In this section, the procedure of applying frequency warping to normalize filterbank-based acoustic features is discussed. Generally, frequency warping in the mel-filterbank is a well-known technique that was utilized in speech recognition tasks for speaker normalization [12]. Sheikhan et al. [15] also used this approach for the normalization of the MFCC feature for the emotional states of *Anger* and *Happy* in EASR. This strategy is also adopted in the present work for the normalization of other filterbank-based features for different emotional states (see Fig. 2).

Based on this strategy, frequency warping is applied to the frequencies of a typical filterbank to change the positions of frequency components. In this work, the distinction between vocal tract length of the emotional and neutral speech is modeled by a linear frequency warping function. The warping is performed by a piecewise linear function to preserve the bandwidth of the original signal. Motivated by the approach introduced in [12], the following warping function is proposed to perform the frequency warping in the filterbank stage of extracting acoustic features:

$$f_{\text{warped}}(n) = \begin{cases} f(n) & f \le f_{2l} \\ \alpha(f(n) - f_{2l}) + f_{2l} & f_{2l} \le f \le f_{2h} \\ \dfrac{(f_{3h} - f_{2l}) - \alpha(f_{2h} - f_{2l})}{f_{3h} - f_{2h}}(f(n) - f_{3h}) + f_{3h} & f_{2h} \le f \le f_{3h} \\ f(n) & f \ge f_{3h}. \end{cases} \qquad (2)$$

In this equation, *f*(*n*) are the frequency bins of the n^{th} frame, $f_{\text{warped}}(n)$ are the corresponding warped frequencies, and the parameter α is the warping factor that

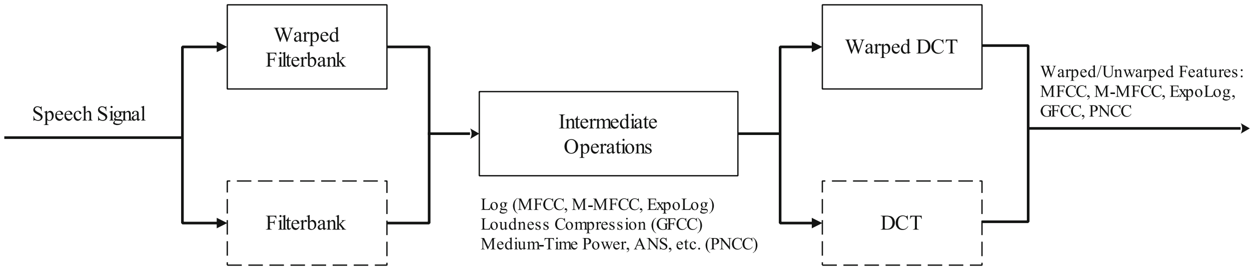

Speech Signal
Warped Filterbank
Filterbank
Intermediate Operations
Log (MFCC, M-MFCC, ExpoLog)
Loudness Compression (GFCC)
Medium-Time Power, ANS, etc. (PNCC)
Warped DCT
DCT
Warped/Unwarped Features:
MFCC, M-MFCC, ExpoLog,
GFCC, PNCC

Fig. 2 The process of frequency warping in the filterbank and/or DCT domain(s) to obtain the corresponding warped features of MFCC, M-MFCC, ExpoLog, GFCC, and PNCC in one or both of the warping domains. The dashed boxes represent the no-frequency warping cases. Intermediate operations performed for each feature extraction method are also shown

controls the amount of warping. Here, formant frequencies are considered to determine the warping intervals, where f_{2l} and f_{2h} represent, respectively, the lowest and highest values of second formants, and f_{3h} depicts the highest value of third formants. These values are obtained as the average values among all second and third formants extracted from the whole sentences of a particular emotional state. The warping factor α for a specific emotional state is computed as the ratio of the average value of the second formants obtained from neutral utterances to that obtained from emotional utterances. The frequency warping is performed in the range of (f_{2l}, f_{2h}), and the linear transformation in the (f_{2h}, f_{3h}) gap is utilized to compensate the spectral changes caused by the frequency warping and return the warping factor to 1. As an example, Fig. 3 shows the distribution of formant frequencies obtained from all utterances of the male speaker in the Persian ESD database [17] for the emotional state of *Disgust* along with the values for the warping intervals and warping factor. Figure 4 illustrates the piecewise linear warping function obtained for a sample utterance of *Disgust* in the database using Eq. (2). Here, the horizontal axis represents the unwarped (i.e., emotional) frequencies whereas the vertical axis indicates the warped (i.e., neutral) frequencies.

2.3.1.2 Frequency warping in the DCT domain Here, the procedure of applying frequency warping is examined in the DCT domain to normalize acoustic features. The frequency warping in DCT was employed in speech recognition tasks for speaker normalization [13]. The same approach was utilized by Sheikhan et al. [15] for the normalization of MFCCs extracted from the emotional utterances of *Anger* and *Happy* in EASR. The approach is also adopted in the present work for the normalization of other features in different emotional states.

Referring to Fig. 2, after the processing performed in the units of “Filterbank” and “Intermediate Operations,” the DCT operation is applied to the input signal $\mathbf{L}$ to compute the cepstral coefficients as:

$$\mathbf{c} = \mathbf{C}.\mathbf{L}, \tag{3}$$

where $\mathbf{C}$ is the DCT matrix with the components given as:

$$C_{km} = \left[\alpha_k \cos\left(\frac{\pi(2m-1)k}{2M}\right)\right]_{\substack{0\le k\le N-1 \\ 1\le m\le M}} \tag{4}$$

Here, M represents the number of filters in the filterbank, N is the number of cepstral coefficients, and α_k is a factor calculated as:

$$\alpha_k = \begin{cases} \sqrt{\dfrac{1}{M}}, & (k=0) \\ \sqrt{\dfrac{2}{M}}. & (k=1,2,\ldots,N-1) \end{cases} \tag{5}$$

In the following, the linear frequency warping in the DCT domain is described for those features that have the DCT calculation in their extraction process [13].

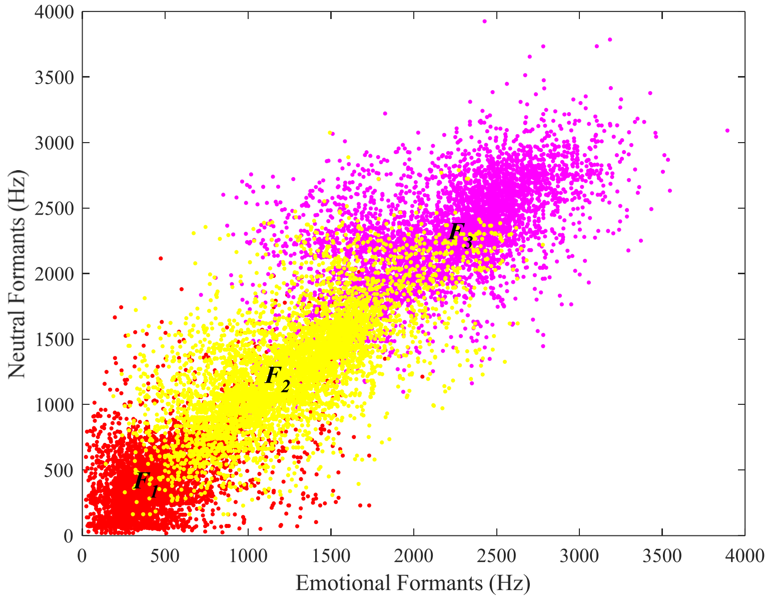


Fig. 3 The distribution of formant frequencies, F_1, F_2, and F_3, for *Disgust* obtained from all utterances of the male speaker in the Persian ESD database [17]

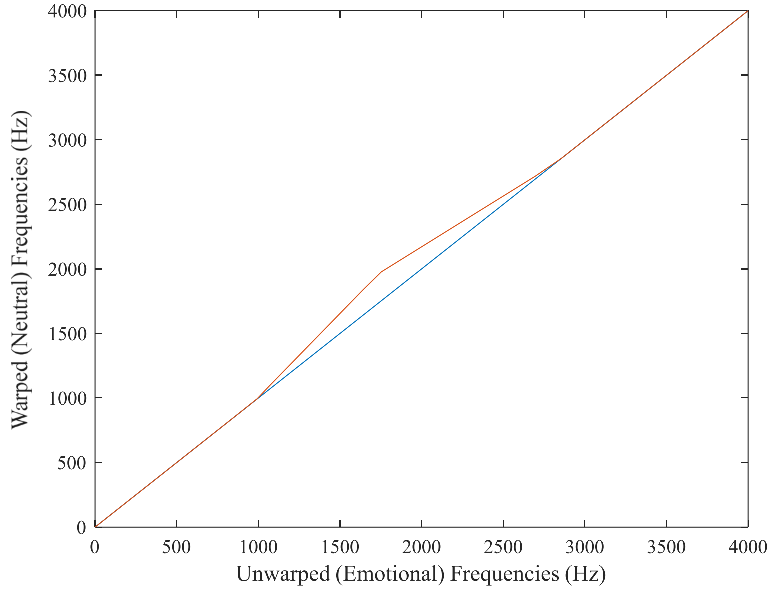


Fig. 4 The piecewise linear frequency warping function obtained for a sample utterance of *Disgust* in the Persian ESD database [17], with the warping factor $\alpha = 1.3$, $f_{2l} = 982$, $f_{2h} = 1739$, and $f_{3h} = 2800$

Step 1: The signal **L** is retrieved from the cepstral coefficients using the inverse DCT (IDCT) operator:

$$\mathbf{L} = \mathbf{C}^{-1}.\mathbf{c}. \tag{6}$$

Here, we consider the unitary type-2 DCT matrix for which $\mathbf{C}^{-1} = \mathbf{C}^{T}$. With this assumption **L** can be written in the expanded form as:

$$\mathbf{L}(m) = \sum_{k=0}^{N-1} c(k)\ \alpha_k\ \cos\left(\frac{\pi(2m-1)k}{2M}\right), \quad (m = 1, 2, \ldots, M) \tag{7}$$

where $c(k)$ $(k = 0, 1, \ldots, N - 1)$ are the cepstral coefficients.

Step 2: Considering $\psi(u)$ as the warping function of the continuous variable u, the warped discrete output is obtained by:

$$\begin{aligned}\hat{\mathbf{L}}(m) &= \mathbf{L}(\psi(u))|_{u=m}, \quad (m = 1, 2, \ldots, M)\\ &= \sum_{k=0}^{N-1} c(k)\ \alpha_k\ \cos\left(\frac{\pi(2\psi(m)-1)k}{2M}\right).\end{aligned} \tag{8}$$

The warping function $\psi(u)$ is computed as:

$$\psi(u) = \frac{1}{2} + M.\theta_p\left(\frac{u-1/2}{M}\right), \tag{9}$$

where $\theta_p(\lambda)$ is the normalized frequency warping function given as:

$$\theta_p(\lambda) = \begin{cases} p\lambda, & (0 \le \lambda \le \lambda_0)\\ p\lambda_0 + \left(\frac{1-p\lambda_0}{1-\lambda_0}\right)(\lambda-\lambda_0). & (\lambda_0 \le \lambda \le 1)\end{cases} \tag{10}$$

Here, λ represents the normalized frequency, λ_0 is the normalized reference frequency specifying the range (0, λ_0) in which frequency warping is performed, and p is the warping factor that controls the amount of warping.

By rewriting Eq. (8) in vector form, we obtain:

$$\hat{\mathbf{L}} = \tilde{\mathbf{C}}.\mathbf{c}, \tag{11}$$

where $\tilde{C}$ represents the warped IDCT matrix given as:

$$\tilde{C}_{m,k} = \left[\alpha_k\ \cos\left(\frac{\pi(2\psi(m)-1)k}{2M}\right)\right]_{\substack{1\le m\le M\\ 0\le k\le N-1}} \tag{12}$$

By rearranging Eq. (9), the warped IDCT matrix can be written in terms of normalized frequency warping function $\theta_p(\lambda)$:

$$\frac{2\psi(u)-1}{2M} = \theta_p\left(\frac{2u-1}{2M}\right), \tag{13}$$

$$\tilde{C}_{m,k} = \left[\alpha_k\ \cos\left(\pi k\theta_p\left(\frac{2m-1}{2M}\right)\right)\right]_{\substack{1\le m\le M\\ 0\le k\le N-1}} \tag{14}$$

Step 3: Finally, by putting the warped discrete output $\hat{\mathbf{L}}$ in Eq. (3), the warped cepstral coefficients $\hat{\mathbf{c}}$ are computed as:

$$\begin{aligned}\hat{\mathbf{c}} &= \mathbf{C}.\hat{\mathbf{L}} = \left(\mathbf{C}.\tilde{\mathbf{C}}\right)\mathbf{c},\\ &= \mathbf{T}.\mathbf{c},\end{aligned} \tag{15}$$

where the matrix $\mathbf{T}(\mathbf{T}{=}\mathbf{C}.\tilde{\mathbf{C}})$ is a linear transformation that transforms the initial cepstral coefficients into the warped coefficients.

In the present work, the above approach is applied to acoustical features to obtain the DCT-warped MFCC, M-MFCC, ExpoLog, GFCC, and PNCC. An example of the warping function employed in the DCT unit is depicted in Fig. 5.

Notably, the warping factors p used in the DCT warping for different emotions are calculated in the same manner as α obtained for the filterbank warping (refer to Eq. (2)).

2.3.2 Applying VTLN to acoustic features

In this section, based on the model for the extraction of various acoustical features, different VTLN warping methods are employed in the filterbank and/or DCT domain(s) to obtain warped (i.e., normalized) features which are finally fed into the Kaldi ASR system. To this aim, the filterbank warping is implemented by employing the warping function given in Eq. (2), whereas, in the DCT warping procedure, the steps given in *Frequency warping in the DCT domain* are adopted. The combined frequency warping is obtained by concatenating the respective frequency warping operations in both filterbank and DCT domains.

3 Experiments and evaluations

3.1 Experimental setup

To examine the effectiveness of the frequency warping for MFCC, M-MFCC, ExpoLog, GFCC, and PNCC in the speech recognition system, the performances of these features and their corresponding warped features are evaluated in the Kaldi baseline ASR system for different emotional states.

To this aim, the Persian ESD [17] and CREMA-D [18] datasets are used to train and test the GMM-HMM/DNN-HMM Kaldi speech recognizer.

The baseline system is trained using MFCC, M-MFCC, ExpoLog, GFCC, and PNCC extracted from neutral utterances of databases. The extracted features have all 13 dimensions, except for GFCC which is 23-dimensional. The delta and delta-delta features are also calculated and added to the previously extracted features to construct a complete acoustic feature. The training procedure of the Kaldi baseline consists of constructing appropriate lexicons, generating language models, and training the acoustic models of the corresponding databases. First, the lexicons of the Persian ESD and CREMA-D are generated. Then, the corresponding language models are produced according to the constructed lexicons. In the training of the acoustic models in the GMM-HMM-based system, 3-state monophone HMMs are used to model all

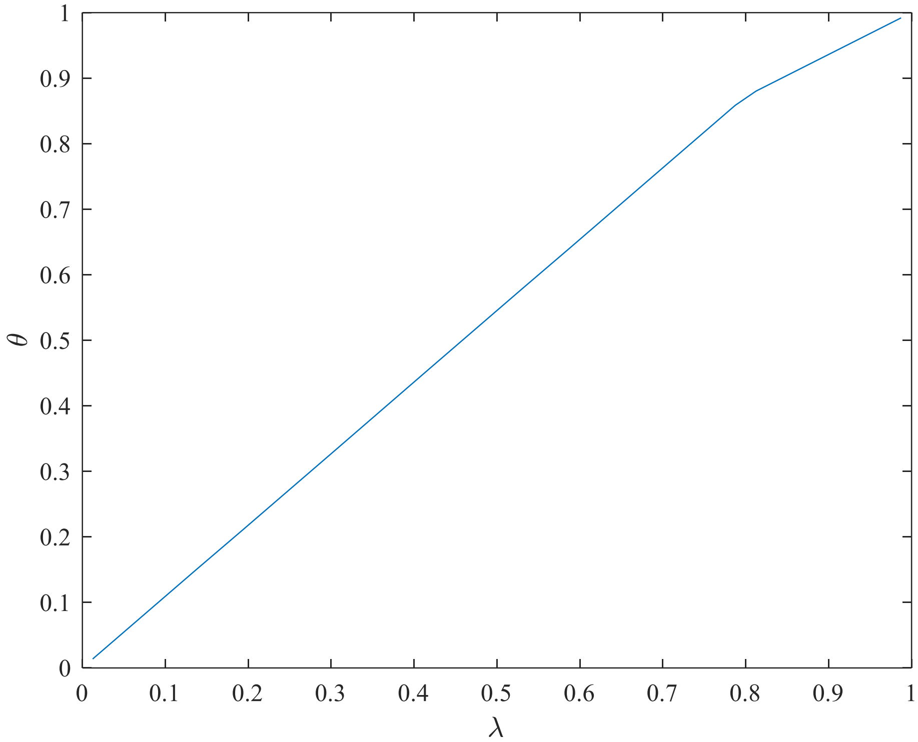


Fig. 5 Warping function based on Eq. (14) plotted for the emotional state of *Happy* in the Persian ESD database [17] with the warping factor $p = 1.09$, and $\lambda_0 = 0.8$

phonemes in the datasets (30 in Persian ESD, 38 in CREMA-D), including silences and pauses. In contrast, the training of the acoustic models in the DNN-HMM-based system requires triphone HMMs. In this paper, the training of the DNN-HMM EASR system is performed based on Karel Vesely's method [30] in the Kaldi toolkit.

The performance of the proposed EASR system is assessed in three experiments. In the first experiment, the effectiveness of CMN [12] is inspected without employing the VTLN method in the EASR system. Here, first, the GMM-HMM-based system is trained with/without employing the CMN technique in extracting features from neutral utterances. Then, speech recognition is performed based on the features extracted both from the emotional and neutral utterances of the corpora.

In the second experiment, the impact of employing different values of the normalized reference frequency, λ_0, is studied on the performance of the GMM-HMM Kaldi for different acoustic features. The optimal λ_0 is then chosen for the later frequency warping experiments.

In the last experiment, the advantage of using warped emotional speech features in the GMM-HMM/DNN-HMM speech recognition system is explored. Here, the simulations are conducted with different structures of the Kaldi speech recognizer. First, both the Persian ESD and CREMA-D datasets are used to train and test the GMM-HMM Kaldi system. Then, CREMA-D with sufficient utterances and speakers is employed to evaluate the recognition performance of the warped features with the state-of-the-art DNN-HMM Kaldi recognizer. By considering the benefits of employing the CMN technique in the EASR system as observed in the first experiment, the CMN procedure is applied to all features to compensate for speaker variability in the Kaldi system. Here, the performances of warped features in the filterbank and/or DCT domain(s) are compared with those of unwarped features in the Kaldi baseline.

The evaluation experiments of the proposed EASR system are conducted for five emotional states, including *Anger*, *Disgust*, *Fear*, *Happy*, and *Sad*.

3.2 Databases

The experimental evaluations are carried out by the Persian emotional speech database (Persian ESD) [17] and crowd-sourced emotional multi-modal actors dataset (CREMA-D) [18]. Table 1 illustrates briefly the specifications of the Persian ESD and CREMA-D databases.

The Persian ESD is a script-fixed dataset that encompasses comprehensive emotional speech of standard Persian language containing a validated set of 90 sentences. These sentences were uttered in different emotional states (i.e., *Anger*, *Disgust*, *Fear*, *Happy*, and *Sad*) and neutral mode by two native Persian speakers (one male and one female). The recording of the database was accomplished in a professional recording studio in Berlin under the supervision of acoustic experts. As shown in Table 1, the Persian ESD comprises 472 speech utterances, each with a duration of 5 s on average, which are classified into five aforementioned basic emotional groups. The database was articulated in three situations: (1) congruent: emotional lexical content spoken in a congruent emotional voice (76 sentences by two speakers), (2) incongruent: neutral sentences spoken in an emotional voice (70 sentences by two speakers), and (3) baseline: all emotional and neutral sentences spoken in neutral voice (90 sentences by two speakers). In general, sentences with different emotions do not have the same lexical content. The validity of the database was evaluated by a group of 34 native speakers in a perception test. Utterances having a recognition rate of 71.4% or better were regarded as valid descriptions of the target emotions. The recordings are available at a sampling rate of 44.1 kHz and mono channel.

The CREMA-D is an audio-visual script-fixed dataset for the study of multi-modal expression and perception of basic acted emotions. As shown in Table 1, the dataset consists of a collection of 7442 original clips of 91 actors (48 males, 43 females, age 20–74) of various races and ethnicities with facial and vocal emotional expressions in sentences. The actors uttered 12 sentences with various emotional states (*Anger*, *Disgust*, *Fear*, *Happy*, *Neutral*, and *Sad*) and different emotion levels (Low, Medium, High, and Unspecified). Sentences with different emotions have the same lexical content. Using perceptual ratings from crowd sourcing, the database was submitted for validation by 2443 raters to evaluate the categorical emotion labels and real-value intensity values for the perceived emotion. Participants assessed the dataset in audio-only, visual-only, and audio-visual modalities with recognition rates of 40.9%, 58.2%, and 63.6%, respectively. In this paper, audio-only data with a sampling rate of 16 kHz are used.

3.3 Evaluation criterion

The performance of an ASR system for a particular task is often measured by comparing the hypothesized and test transcriptions. In this context, the percentage of word error rate (WER), as the most widely used metric, is used to evaluate the recognition performance of the proposed EASR system. After the alignment of the two-word sequences (i.e., hypothesis and test), WER is calculated by the rate of the number of errors to the total number of words in the test utterances.

Table 1 The specifications of the Persian ESD and CREMA-D databases

		Persian ESD			CREMA-D		
		Male	**Female**	**Total**	**Male**	**Female**	**Total**
Number of utterances	Anger	31	31	62	671	600	1271
	Disgust	29	29	58	671	600	1271
	Fear	29	29	58	671	600	1271
	Happy	29	29	58	671	600	1271
	Neutral	90	90	180	575	512	1087
	Sad	28	28	56	671	600	1271
	Total	236	236	472	3930	3512	7442
Number of speakers		1	1	2	48	43	91
Number of sentences				90			12
Sampling rate				44.1 kHz			16 kHz

3.4 Results and discussions

3.4.1 Effect of CMN on the recognition performance

In the first experiment, the effect of employing CMN in the EASR system is explored. For this purpose, the baseline ASR system trained with neutral utterances is tested with emotional speech inputs (i.e., unmatched scenario) with/without applying CMN to the extracted features. For comparisons, the performance of the baseline system is also evaluated in the case of neutrally trained/neutrally tested inputs (i.e., matched scenario). For the simulations of this scenario, the neutral utterances are split into two sets; 80% for training and the remaining 20% for testing. The experimental results are shown in Figs. 6 and 7 for the Persian ESD [17] and CREMA-D [18] datasets, respectively.

The outcomes of this experiment for both databases reveal clearly that using the CMN method for the unmatched scenario yields, on average, superior performance of the recognizer in terms of WER. This implies that CMN decreases the destructive effects of emotional speech. The evaluation results in the case of matched scenario show no considerable effect of CMN on the recognizer efficiency. Also, the recognition results in Figs. 6 and 7 show significantly different recognition performances for the emotional and neutral input utterances. It is obvious that in the case of the neutrally trained/neutrally tested experiment, the WER values are small for the baseline system trained with different features. However, when the neutrally trained system is tested with emotional utterances, in general, WERs are increased extremely. This fact indicates that the emotion-affected speech represents a significant mismatch condition of the ASR systems trained with neutral utterances. Furthermore, by comparing the average WER scores among all emotional states obtained for both databases, one realizes that PNCC and GFCC have the best and worst performance, respectively, introducing PNCC as the robust feature in the EASR system. Due to the benefits of employing CMN in decreasing WER, in the following experiments, we use the CMN technique in the construction of the features.

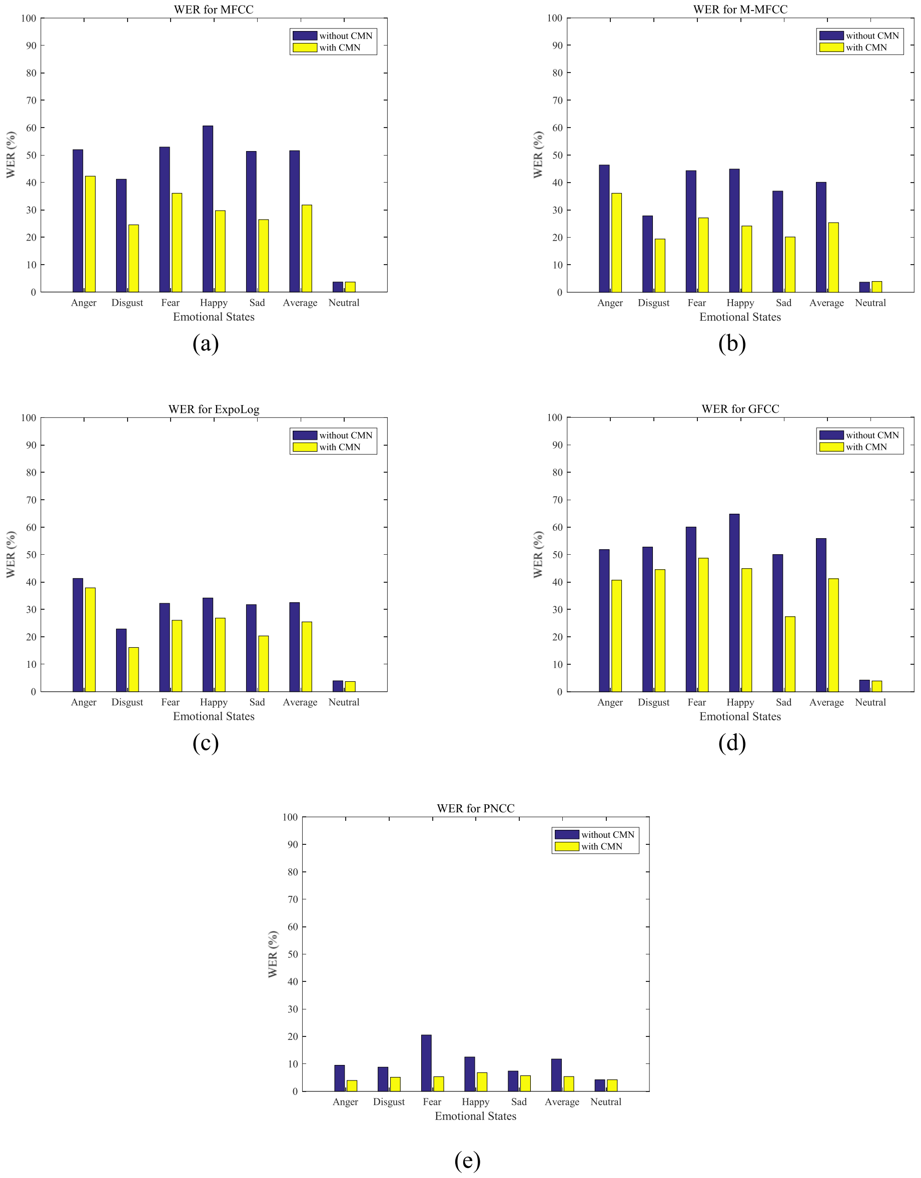
3.4.2 Investigation of $\lambda_0$values in frequency warping

Here, the impact of the normalized reference frequency, λ_0, on the recognition performance of different acoustic features in the DCT domain and the combined filterbank and DCT domain is examined. The results of this analysis for three values of λ_0 are represented in Tables 2 and 3, respectively, for the Persian ESD and CREMA-D datasets. Here, among all features investigated, it can be observed that PNCC has the highest performance in the GMM-HMM EASR system due to its robustness against different emotional states. Also, changing the value of λ_0 has no sensible impact on improving the recognition results of this feature. By comparing the average WER values shown in Tables 2 and 3, it can be seen (except for ExpoLog in “DCT Warping” and ExpoLog and GFCC in “Filterbank & DCT Warping”) that the best performance is achieved by $\lambda_0 = 0.4$ for all acoustic features. This value of λ_0 is considered in the following experiments to specify the range of frequency warping.

3.4.3 Frequency warping in the EASR system

The last experiment concerns evaluating the efficiency of frequency warping in the filterbank and/or DCT domain(s) for the neutrally trained/emotionally tested GMM-HMM/DNN-HMM EASR system.

3.4.3.1 GMM-HMM EASR system Tables 4 and 5 represent the performance scores of the GMM-HMM EASR system for the warped emotional features of MFCC, M-MFCC, ExpoLog, GFCC, and PNCC in the filterbank and/or DCT domain(s) as compared with those of the unwarped features for the Persian ESD [17] and CREMA-D [18] datasets, respectively. Comparing the results of both tables shows, in general, that the WER



(a)

(b)

(c)

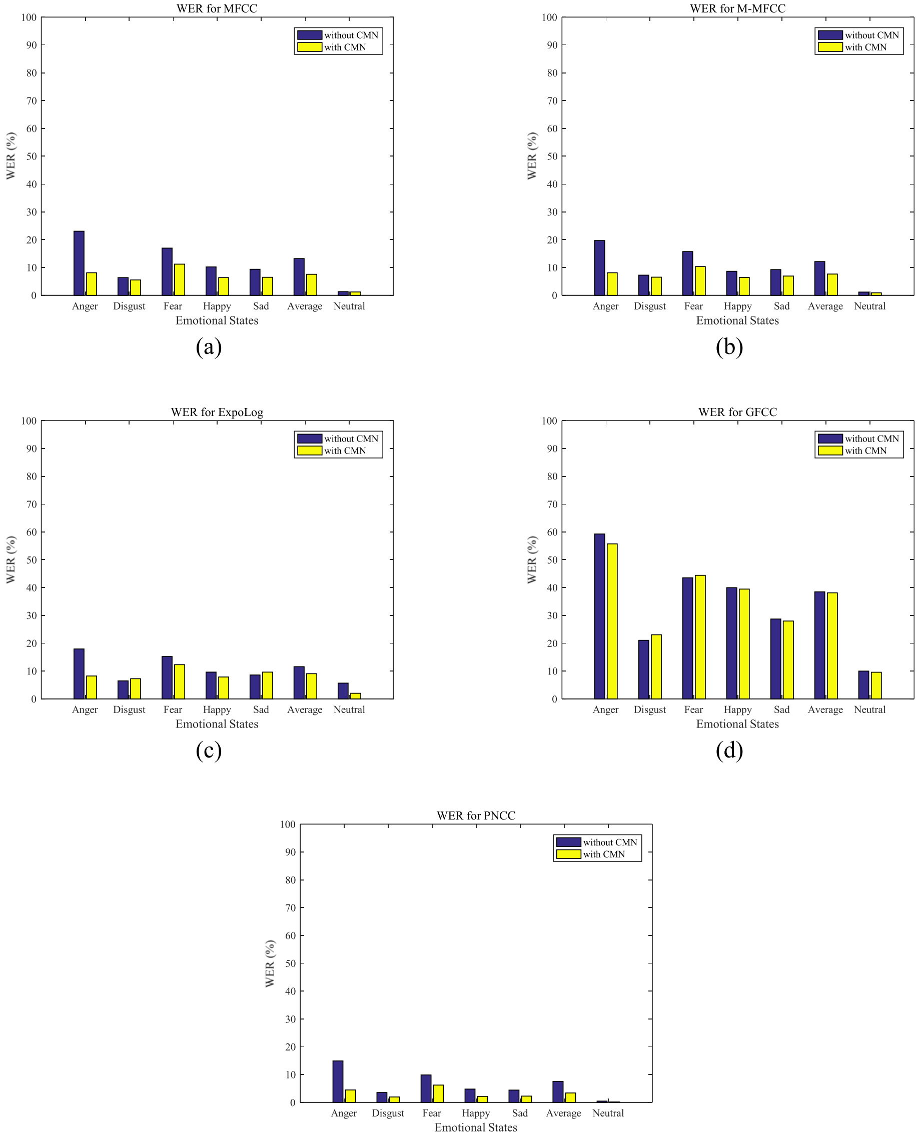
(d)

(e)

Fig. 6 WER scores (%) for the Persian ESD dataset with/without applying the CMN method to the acoustic features of **a** MFCC, **b** M-MFCC, **c** ExpoLog, **d** GFCC, and **e** PNCC

values for the CREMA-D database are lower than the corresponding values for the Persian ESD dataset. This observation can be justified by the fact that the number of neutral utterances used in CREMA-D to train the GMM-HMM EASR is much higher than those given in Persian ESD (1087 utterances vs. 180 utterances).

The evaluations presented in the tables can be interpreted from two perspectives; from the aspect of the applied warping methods, and the aspect of the acoustic features used in the construction of the EASR system. By applying different warping methods to the acoustic features, the results of both tables show that employing all variants of the frequency warping methods improves the recognition rates in terms of WER. In the case of PNCC, WER values are close to each other for all warping methods, showing no advantage of any warping technique over others. Also, by comparing the average values of WER, it is observed, in general, that (except for ExpoLog) the effectiveness of applying the DCT warping to the features is more superior to the filterbank warping and the combined filterbank and DCT warping procedure. Considering the success of applying the DCT warping in decreasing the destructive effect of emotion in the EASR system, this can be interpreted as saying that no further improvement is reached by adding the capability of filterbank warping to the DCT normalization process.



(a)

(b)

(c)

(d)

(e)

Fig. 7 WER scores (%) for the CREMA-D dataset with/without applying the CMN method to the acoustic features of **a** MFCC, **b** M-MFCC, **c** ExpoLog, **d** GFCC, and **e** PNCC

The results given in the tables can also be interpreted based on the performances of different acoustical features used in the implementation of the EASR system. Comparing the average WER scores obtained for various warped features in all emotional states indicates that PNCC attains the lowest WER score, whereas GFCC achieves the highest score. Accordingly, among different acoustical features, the warped PNCC can be employed as a robust feature in the EASR system. This confirms the results obtained for PNCC in the first experiment concerning the benefits of applying CMN to the features. The high performance of PNCC is associated with the use of different processing stages in the implementation of PNCC, including the use of a medium-time processing, power-law nonlinearity, a noise suppression algorithm based on asymmetric filtering, and a module that accomplishes temporal masking [24]. Especially, in the medium-time processing of PNCC, longer analysis window frames are considered, which are proved to provide better performance for noise modeling and/or environmental normalization. The use of PNCC has been verified successfully in emotion recognition [31] and emotional speech recognition [32] tasks.

Table 2 The effect of modifying the normalized reference frequency, λ_0, on the recognition performance of the proposed GMM-HMM EASR system (in terms of WER (%)) for Persian ESD. The values of WER are obtained by applying different warping methods to various acoustic features extracted from different emotional utterances

Feature type	**Warping type**		**Emotional states**					
			Anger	**Disgust**	**Fear**	**Happy**	**Sad**	**Average WER**
MFCC	DCT Warping	$\lambda_0 = 0$	42.30	24.54	36.08	29.70	26.43	31.81
		$\lambda_0 = 0.4$	28.20	22.34	31.32	18.78	18.25	23.78
		$\lambda_0 = 0.7$	38.36	21.79	31.87	19.14	21.48	26.53
	Filterbank & DCT Warping	$\lambda_0 = 0$	42.13	23.81	34.25	30.05	21.48	30.34
		$\lambda_0 = 0.4$	27.05	21.61	32.42	20.21	15.97	23.45
		$\lambda_0 = 0.7$	40.82	22.71	33.52	22.72	21.29	28.21
M-MFCC	DCT Warping	$\lambda_0 = 0$	36.07	19.41	27.11	24.15	20.15	25.38
		$\lambda_0 = 0.4$	17.21	16.30	21.61	16.10	19.01	18.05
		$\lambda_0 = 0.7$	29.67	15.93	21.98	17.35	17.49	20.48
	Filterbank & DCT Warping	$\lambda_0 = 0$	34.26	17.95	25.82	23.79	20.72	24.51
		$\lambda_0 = 0.4$	22.79	15.93	24.73	17.35	20.53	20.27
		$\lambda_0 = 0.7$	31.48	15.38	27.84	18.96	18.25	22.38
ExpoLog	DCT Warping	$\lambda_0 = 0$	37.87	16.12	26.01	26.83	20.34	25.43
		$\lambda_0 = 0.4$	37.54	14.65	25.64	27.55	20.34	25.13
		$\lambda_0 = 0.7$	30.00	13.00	23.26	20.04	16.54	20.57
	Filterbank & DCT Warping	$\lambda_0 = 0$	30.49	13.55	15.02	22.72	15.78	19.51
		$\lambda_0 = 0.4$	32.46	13.92	16.12	28.98	19.01	22.10
		$\lambda_0 = 0.7$	26.89	12.82	17.22	22.18	20.53	19.92
GFCC	DCT Warping	$\lambda_0 = 0$	40.66	44.51	48.72	44.90	27.38	41.23
		$\lambda_0 = 0.4$	26.89	39.74	43.77	39.53	26.81	35.35
		$\lambda_0 = 0.7$	28.52	40.66	43.96	42.58	31.94	37.53
	Filterbank & DCT Warping	$\lambda_0 = 0$	39.67	43.77	47.44	44.01	27.19	40.42
		$\lambda_0 = 0.4$	28.36	40.84	45.60	40.79	29.09	36.94
		$\lambda_0 = 0.7$	30.82	40.84	44.51	43.83	34.79	30.96
PNCC	DCT Warping	$\lambda_0 = 0$	3.93	5.13	5.31	6.80	5.70	5.37
		$\lambda_0 = 0.4$	4.10	5.13	4.40	6.26	4.56	4.89
		$\lambda_0 = 0.7$	4.10	5.13	4.40	6.26	4.56	4.89
	Filterbank & DCT Warping	$\lambda_0 = 0$	3.93	5.13	4.58	6.62	5.70	5.19
		$\lambda_0 = 0.4$	3.93	5.31	4.95	6.44	5.32	5.19
		$\lambda_0 = 0.7$	3.93	5.31	4.95	6.44	5.32	5.19

To recognize the importance of the warping methods as compared with "No Warping" for each feature and emotional state, the *t* test [33] is employed as a statistical analysis tool. The symbol * in Tables 4 and 5 indicates the significant cases (i.e., p value < 0.05). According to the results of the statistical analysis, except *Disgust*, significant values of the WER are observed in most cases of warping methods applied to the corresponding emotional states.

3.4.3.2 DNN-HMM EASR system Speech recognition systems employ HMMs to deal with speech temporal variations. Generally, such systems use GMMs to determine how each state of an HMM fits a frame or a window of frames of coefficients representing the acoustic input. A feed-forward neural network is an alternative way to estimate the fit. This neural network takes several frames of coefficients and generates posterior probabilities over HMM states. Research on speech recognizers shows that the use of DNN in acoustic modeling outperforms the traditional GMM on a variety of databases [34, 35]. This is partly due to the accurate estimation of the state-specific probabilities and better distinguishing of the

Table 3 The effect of modifying the normalized reference frequency, λ_0, on the recognition performance of the proposed GMM-HMM EASR system (in terms of WER (%)) for CREMA-D. The values of WER are obtained by applying different warping methods to various acoustic features extracted from different emotional utterances

Feature type	**Warping type**		**Emotional states**					
			Anger	**Disgust**	**Fear**	**Happy**	**Sad**	**Average WER**
MFCC	DCT Warping	$\lambda_0 = 0$	8.11	5.54	11.20	6.39	6.50	7.55
		$\lambda_0 = 0.4$	7.78	5.47	9.20	4.92	5.71	6.62
		$\lambda_0 = 0.7$	7.19	5.74	10.06	4.98	6.18	6.83
	Filterbank & DCT Warping	$\lambda_0 = 0$	7.80	5.09	9.69	5.40	6.06	6.81
		$\lambda_0 = 0.4$	7.78	5.44	9.25	4.67	5.65	6.56
		$\lambda_0 = 0.7$	7.60	6.70	10.91	5.60	7.20	7.60
M-MFCC	DCT Warping	$\lambda_0 = 0$	8.11	6.53	10.33	6.42	6.97	7.67
		$\lambda_0 = 0.4$	7.28	6.35	9.50	5.79	6.09	7.00
		$\lambda_0 = 0.7$	7.67	6.25	9.32	5.84	6.85	7.19
	Filterbank & DCT Warping	$\lambda_0 = 0$	7.38	6.26	9.32	6.06	6.77	7.16
		$\lambda_0 = 0.4$	7.42	6.16	9.32	5.90	6.50	7.06
		$\lambda_0 = 0.7$	8.33	6.85	10.19	6.19	7.88	7.89
ExpoLog	DCT Warping	$\lambda_0 = 0$	8.25	7.25	12.32	7.88	9.63	9.07
		$\lambda_0 = 0.4$	7.47	6.87	11.48	6.41	8.73	8.19
		$\lambda_0 = 0.7$	7.52	7.17	11.38	5.85	8.53	8.09
	Filterbank & DCT Warping	$\lambda_0 = 0$	7.17	6.63	11.22	6.49	8.75	8.05
		$\lambda_0 = 0.4$	6.97	7.00	10.56	6.11	8.52	7.83
		$\lambda_0 = 0.7$	7.22	7.47	10.33	5.94	9.10	8.01
GFCC	DCT Warping	$\lambda_0 = 0$	55.73	23.07	44.39	39.48	27.97	38.13
		$\lambda_0 = 0.4$	55.33	22.68	43.67	35.85	27.97	37.10
		$\lambda_0 = 0.7$	56.64	24.46	45.28	37.09	30.07	38.71
	Filterbank & DCT Warping	$\lambda_0 = 0$	54.95	22.33	43.28	38.36	26.89	37.16
		$\lambda_0 = 0.4$	55.99	23.02	44.31	36.82	29.32	37.89
		$\lambda_0 = 0.7$	57.76	25.49	46.39	38.69	32.09	40.08
PNCC	DCT Warping	$\lambda_0 = 0$	4.49	1.96	6.25	2.17	2.26	3.43
		$\lambda_0 = 0.4$	4.52	2.03	6.75	3.08	2.37	3.75
		$\lambda_0 = 0.7$	4.52	2.03	6.75	3.08	2.37	3.75
	Filterbank & DCT Warping	$\lambda_0 = 0$	3.97	1.83	5.96	1.71	1.99	3.09
		$\lambda_0 = 0.4$	4.44	1.98	6.21	2.56	2.09	3.46
		$\lambda_0 = 0.7$	4.44	1.98	6.21	2.56	2.09	3.46

class boundaries which result in higher state-level classification performance of HMMs. In this section, the performance of frequency warping is examined with the state-of-the-art DNN-HMM Kaldi speech recognizer using CREMA-D as the database.

The results of the emotional speech recognition for the warped features of MFCC, M-MFCC, ExpoLog, GFCC, and PNCC in the filterbank and/or DCT domain(s) are depicted in Table 6 as compared with those of unwarped features for the CREMA-D dataset. The results illustrate, in general, that using all variants of the frequency warping methods increases the recognition performance of the EASR system. A comparison of the average recognition performances for various warped features in all emotional states reveals that PNCC acquires the lowest WER score, whereas GFCC obtains the highest score. Hence, among various acoustical features, the warped PNCC can be considered as a robust feature in the EASR system. These findings are consistent with the results obtained in the GMM-HMM EASR experiments.

As in the GMM-HMM-based system, in this experiment, the statistical analysis tool of *t* test [33] is used for identifying the importance of the various warping

Table 4 The recognition performance of the proposed GMM-HMM EASR system (in terms of WER (%)) for Persian ESD. The values of WER are obtained by applying different warping methods to various acoustic features extracted from different emotional utterances. The symbol * shows statistically significant cases (i.e., p value < 0.05)

Feature type	Warping type	Emotional states					
		Anger	Disgust	Fear	Happy	Sad	Average WER
MFCC	No Warping	42.30	24.54	36.08	29.70	26.43	31.81
	Filterbank Warping	42.13	23.81	34.25	30.05	21.48*	30.34
	DCT Warping	28.20*	22.34	31.32*	18.78*	18.25*	23.78
	Filterbank & DCT Warping	27.05*	21.61*	32.42*	20.21*	15.97*	23.45
M-MFCC	No Warping	36.07	19.41	27.11	24.15	20.15	25.38
	Filterbank Warping	34.26	17.95	25.82*	23.79	20.72	24.51
	DCT Warping	17.21*	16.30	21.61*	16.10*	19.01	18.05
	Filterbank & DCT Warping	22.79*	15.93	24.73*	17.35*	20.53	20.27
ExpoLog	No Warping	37.87	16.12	26.01	26.83	20.34	25.43
	Filterbank Warping	30.49*	13.55*	15.02*	22.72*	15.78*	19.51
	DCT Warping	37.54	14.65	25.64	27.55	20.34*	25.13
	Filterbank & DCT Warping	32.46*	13.92*	16.12*	28.98	19.01*	22.10
GFCC	No Warping	40.66	44.51	48.72	44.90	27.38	41.23
	Filterbank Warping	39.67	43.77	47.44	44.01	27.19*	40.42
	DCT Warping	26.89*	39.74*	43.77	39.53	26.81	35.35
	Filterbank & DCT Warping	28.36*	40.84*	45.60	40.79	29.09*	36.94
PNCC	No Warping	3.93	5.13	5.31	6.80	5.70	5.37
	Filterbank Warping	3.93	5.13	4.58	6.62	5.70	5.19
	DCT Warping	4.10	5.13	4.40*	6.26	4.56	4.89
	Filterbank & DCT Warping	3.93*	5.31*	4.95	6.44	5.32*	5.19

methods in comparison with "No Warping" for each feature and emotional state. The significant cases of the test (i.e., p value < 0.05) are specified with the symbol * in Table 6.

Comparing the cases of "No Warping" for different features and emotional states in Tables 5 and 6, it can be observed that the DNN-HMM EASR system outperforms the GMM-HMM-based system in terms of the WER values. This could be expected, since, as it was explained before, DNN has a higher performance than the traditional GMM as to the acoustic modeling in speech recognition systems. Also, comparing the WER values in both tables shows that the number of significant cases in Table 6 is lower than that in Table 5. This again can be justified by the fact that the DNN-HMM Kaldi has a better performance than the GMM-HMM Kaldi which prevents the warping methods from having a large impact in reducing the mismatch between the neutral training and emotional testing conditions. However, in contrast with the GMM-HMM Kaldi system, the DNN-HMM Kaldi speech recognizer requires a larger database and more computational time and complexity in training/testing phases.

3.4.3.3 Comparisons between different frequency warping methods As a further evaluation process, a comparison has been performed between the results obtained by the proposed GMM-HMM/DNN-HMM EASR system with Persian ESD and CREMA-D as databases and those obtained by Sheikhan et al. [15]. In this context, it is noteworthy that the experiments conducted by Sheikhan et al. [15] were limited only to MFCC as the feature and *Anger* and *Happy* as the emotions. In contrast, our simulations consider more emotional states and acoustic features. Table 7 gives a summary of the performance comparisons between different warping methods for the specified features and emotions, where the symbol ">" is interpreted as "better" and "≯" means "not better."

4 Conclusion

In this paper, the improvement of the ASR system for emotional input utterances is investigated, where the mismatch between training and recognition conditions results in a significant reduction in the performance of the system. The main objective of the proposed EASR

Table 5 The recognition performance of the proposed GMM-HMM EASR system (in terms of WER (%)) for CREMA-D. The values of WER are obtained by applying different warping methods to various acoustic features extracted from different emotional utterances. The symbol * shows statistically significant cases (i.e., *p* value < 0.05).

Feature type	Warping type	Emotional states					
		Anger	Disgust	Fear	Happy	Sad	Average WER
MFCC	No Warping	8.11	5.54	11.20	6.39	6.50	7.55
	Filterbank Warping	7.80	5.09	9.69*	5.40*	6.06	6.81
	DCT Warping	7.78*	5.47	9.20*	4.92*	5.71*	6.62
	Filterbank & DCT Warping	7.78*	5.44	9.25*	4.67*	5.65*	6.56
M-MFCC	No Warping	8.11	6.53	10.33	6.42	6.97	7.67
	Filterbank Warping	7.38*	6.26	9.32*	6.06	6.77	7.16
	DCT Warping	7.28	6.35	9.50*	5.79	6.09*	7.00
	Filterbank & DCT Warping	7.42*	6.16	9.32*	5.90	6.50	7.06
ExpoLog	No Warping	8.25	7.25	12.32	7.88	9.63	9.07
	Filterbank Warping	7.17*	6.63	11.22*	6.49*	8.75	8.05
	DCT Warping	7.47	6.87	11.48*	6.41*	8.73	8.19
	Filterbank & DCT Warping	6.97*	7.00	10.56*	6.11*	8.52*	7.83
GFCC	No Warping	55.73	23.07	44.39	39.48	27.97	38.13
	Filterbank Warping	54.95	22.33	43.28*	38.36*	26.89*	37.16
	DCT Warping	55.33	22.68	43.67*	35.85*	27.97*	37.10
	Filterbank & DCT Warping	55.99	23.02	44.31	36.82*	29.32	37.89
PNCC	No Warping	4.49	1.96	6.25	2.17	2.26	3.43
	Filterbank Warping	3.97*	1.83	5.96	1.71*	1.99*	3.09
	DCT Warping	4.52	2.03	6.75	3.08	2.37	3.75
	Filterbank & DCT Warping	4.44	1.98	6.21	2.56	2.09	3.46

system is to mitigate the effects of emotional speech and to enhance the efficiency of the recognition system. For this purpose, the VTLN method is employed in the feature extraction stage to decrease the effects of emotion in the recognition process. This goal is achieved by applying the frequency warping in the filterbank and/or DCT domain(s). Accordingly, it is expected that the performance of the warped emotional features approaches that of the corresponding neutral features. The proposed system incorporates the Kaldi ASR as the back end which is trained with the different acoustical features (i.e., MFCC, M-MFCC, ExpoLog, GFCC, and PNCC) extracted from neutral utterances. The EASR system trained with neutral utterances is tested with emotional speech inputs in emotional states of *Anger*, *Disgust*, *Fear*, *Happy*, and *Sad*. The Persian emotional speech dataset (Persian ESD) and crowd-sourced emotional multimodal actors dataset (CREMA-D) are used for the simulations.

In the experiments, first, the effectiveness of the CMN method is investigated in the recognition performance of the emotional utterances for the neutrally trained/emotionally tested GMM-HMM ASR system. The results of this experiment show that employing this technique improves the recognition scores. Then, the influence of using different values of the normalized reference frequency, λ_0, is inspected on the performance of the GMM-HMM-based system. The results of this experiment lead to the selection of an optimal λ_0 for the later experiments. To evaluate the performance of the proposed EASR system, the last experiment explores the advantage of using warped features in the GMM-HMM/DNN-HMM speech recognition system. It is observed, in general, that employing all variants of the frequency warping methods improves the recognition performance of both EASR systems in terms of WER. Also, the experimental results show that the DNN-HMM EASR system achieves higher performance than the GMM-HMM-based system in reducing the mismatch between the neutral training and emotional testing conditions. The higher performance of the DNN-HMM-based system is due to the use of DNN for acoustic modeling in the structure of Kaldi ASR. A comparison of different warped features in both GMM-HMM and DNN-HMM EASR systems confirms that the best WER score is attained for PNCC, whereas the worst score is achieved

Table 6 The performance of the proposed DNN-HMM EASR system (in terms of WER (%)) for CREMA-D by applying different warping methods to various acoustic features and emotional states. The average WER values are given in the last column. The symbol * shows statistically significant cases (i.e., *p* value < 0.05)

Feature type	**Warping type**	**Emotional states**					
		Anger	**Disgust**	**Fear**	**Happy**	**Sad**	**Average WER**
MFCC	No Warping	3.23	3.34	3.39	1.84	1.58	2.68
	Filterbank Warping	3.54	3.39	4.05*	2.17	1.65	2.96
	DCT Warping	2.91	3.22	3.21	1.74	1.58	2.53
	Filterbank & DCT Warping	3.05	3.29	2.89	1.56	1.46	2.45
M-MFCC	No Warping	3.21	3.54	4.28	1.82	1.99	2.97
	Filterbank Warping	3.36	3.68	4.38	2.28*	2.05	3.15
	DCT Warping	3.08	3.78	3.58*	1.59	1.94	2.79
	Filterbank & DCT Warping	3.10	3.93	3.54*	1.49*	1.77	2.77
ExpoLog	No Warping	2.65	4.03	3.86	1.76	1.77	2.81
	Filterbank Warping	2.33*	3.91	3.73	1.42*	1.78	2.63
	DCT Warping	2.48	3.76	3.96	1.61	1.75	2.71
	Filterbank & DCT Warping	2.30	3.73	3.86	1.52	1.92	2.67
GFCC	No Warping	22.75	5.93	11.67	6.93	3.03	10.06
	Filterbank Warping	22.42	5.79	11.03	6.16*	2.91	9.66
	DCT Warping	25.75*	5.69	9.94*	5.47*	3.35	10.04
	Filterbank & DCT Warping	25.79*	5.73	9.89*	4.97*	3.45*	9.97
PNCC	No Warping	3.31	2.69	2.15	1.37	1.13	2.13
	Filterbank Warping	3.21	2.60	2.27	1.05*	1.04	2.03
	DCT Warping	2.90	2.89	2.40	1.04	1.03	2.05
	Filterbank & DCT Warping	2.65	2.85	2.45	1.19	1.01	2.03

for GFCC. The high performance of PNCC can be justified in the use of different processing stages in the PNCC extraction method, which makes this feature robust against various emotional states.

The focus of this research is based on the normalization of segmental or vocal tract-specific features. However, the speech signal consists of both segmental and supra-segmental (i.e., prosodic) information. It is known that prosodic features such as pitch and intonation can also be influenced by the emotional states of a speaker. As future work, new compensation methods can be devised to normalize such prosodic features together with the vocal tract-related features before feeding them to an ASR system. Furthermore, since emotional speech is generally produced in a real environment, this work can also be extended to operate in scenarios such as reverberant and noisy conditions.

Table 7 The performance comparisons between different frequency warping methods used in the proposed GMM-HMM/DNN-HMM EASR system for the Persian ESD and CREMA-D datasets and the system of Sheikhan et al. [15] for various acoustic features and emotional states

Sheikhan et al. [15]	**Proposed**	
GMM-HMM	**GMM-HMM** **(Persian ESD and CREMA-D)**	**DNN-HMM** **(CREMA-D)**
Features: MFCC Emotions: Anger, Happy	Features: MFCC, M-MFCC, GFCC, and PNCC Emotions: Anger, Disgust, Fear, Happy, and Sad	Features: ExpoLog, GFCC, and PNCC Emotions: Anger, Disgust, Fear, Happy, and Sad
Warping > No Warping	Warping > No Warping	Warping > No Warping
DCT > Filterbank	DCT > Filterbank	DCT ≯ Filterbank
Filterbank & DCT > DCT	Filterbank & DCT ≯ DCT	Filterbank & DCT > DCT
Filterbank & DCT > Filterbank	Filterbank & DCT ≯ Filterbank	Filterbank & DCT ≯ Filterbank

Abbreviations

AM: Acoustic model; ANN: Artificial neural network; ASR: Automatic speech recognition; BNC: British national corpus; CMN: Cepstral mean normalization; CREMA-D: Crowd-sourced emotional multi-modal actors dataset; DCT: Discrete-cosine transform; DNN: Deep neural network; EASR: Emotion-affected speech recognition; ESD: Emotional speech database; ExpoLog: Exponential logarithmic scale; FFCC: F-ratio scale frequency cepstral coefficient; GFCC: Gammatone filterbank cepstral coefficient; GMM: Gaussian mixture model; HMM: Hidden Markov model; IDCT: Inverse discrete-cosine transform; IVR: Interactive voice response; LM: Language model; LPCC: Linear prediction cepstral coefficient; MAP: Maximum a posteriori; MFCC: Mel-frequency cepstral coefficient; MLLR: Maximum likelihood linear regression; M-MFCC: Modified mel-scale cepstral coefficient; MRHMM: Multiple regression hidden Markov model; PNCC: Power normalized cepstral coefficient; PSSZN: Per-speaker-syllable-z-normalization; RASTA-PLP: Relative spectral perceptual linear prediction; RM: Resource management; SER: Sentence error rate; VTLN: Vocal tract length normalization; WER: Word error rate

Acknowledgments

Not applicable.

Authors' contributions

MG, EF, and MB have equally contributed in proposing the ideas, discussing the results, and writing and proofreading the manuscript. EF carried out the implementation and experiments. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The Persian ESD and CREMA-D are not publicly available but can be obtained upon request (refer to references [17, 18]).

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 6 July 2020 Accepted: 29 June 2021
Published online: 04 August 2021

References

1. M. Najafian, Acoustic model selection for recognition of regional accented speech. PhD thesis, University of Birmingham (2016)
2. P.C. Woodland, in *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*. Speaker adaptation for continuous density HMMs: review (2001)
3. B. Vlasenko, D. Prylipko, A. Wendemuth, in *35th German Conference on Artificial Intelligence (KI-2012), Saarbrücken, Germany (September 2012)*. Towards robust spontaneous speech recognition with emotional speech adapted acoustic models (2012), pp. 103–107 Citeseer
4. F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, in *Ninth European Conference on Speech Communication and Technology*. A database of German emotional speech (2005)
5. Y. Pan, M. Xu, L. Liu, P. Jia, in *IMACS Multiconference on Computational Engineering in Systems Applications*. Emotion-detecting based model selection for emotional speech recognition, vol 2 (2006), pp. 2169–2172 IEEE
6. B. Schuller, J. Stadermann, G. Rigoll, in *Proc. Speech Prosody 2006, Dresden*. Affect-robust speech recognition by dynamic emotional adaptation (2006)
7. Y. Ijima, M. Tachibana, T. Nose, T. Kobayashi, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM (IEEE, 2009), pp. 4157–4160
8. T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, C. Cox, ASR for emotional speech: clarifying the issues and enhancing performance. Neural Networks **18**(4), 437–444 (2005)
9. D. Gharavian, M. Sheikhan, M. Janipour, Pitch in emotional speech and emotional speech recognition using pitch frequency. Majlesi J Electrical Eng **4**(1) (2010)
10. D. Gharavian, S. Ahadi, in *the Proceedings of International Symposium on Chinese Spoken Language Processing*. Recognition of emotional speech and speech emotion in Farsi, vol 2 (2006), pp. 299–308 Citeseer
11. Y. Sun, Y. Zhou, Q. Zhao, Y. Yan, in *International Conference on Information Engineering and Computer Science (ICIECS)*. Acoustic feature optimization for emotion affected speech recognition (2009), pp. 1–4 IEEE
12. L. Lee, R. Rose, A frequency warping approach to speaker normalization. IEEE Transactions on Speech and Audio Processing **6**(1), 49–60 (1998)
13. S. Panchapagesan, Frequency warping by linear transformation, and vocal tract inversion for speaker normalization in automatic speech recognition, PhD thesis, University of California, Los Angeles (2008)
14. P. Price, W.M. Fisher, J. Bernstein, D.S. Pallett, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. The DARPA 1000-word resource management database for continuous speech recognition (IEEE, 1988), pp. 651–654
15. M. Sheikhan, D. Gharavian, F. Ashoftedel, Using DTW neural–based MFCC warping to improve emotional speech recognition. Neural Computing and Applications **21**(7), 1765–1773 (2012)
16. D. Povey et al., in *IEEE 2011 workshop on automatic speech recognition and understanding*. The Kaldi speech recognition toolkit (2011) no. EPFL-CONF-192584: IEEE Signal Processing Society
17. N. Keshtiari, M. Kuhlmann, M. Eslami, G. Klann-Delius, Recognizing emotional speech in Persian: a validated database of Persian emotional speech (Persian ESD). Behavior Research Methods **47**(1), 275–294 (2015)
18. H. Cao, D.G. Cooper, M.K. Keutmann, R.C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing **5**(4), 377–390 (2014)
19. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(4), 357–366 (1980)
20. 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**(4), 429–442 (2000)
21. Y. Shao, Z. Jin, D. Wang, S. Srinivasan, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. An auditory-based feature for robust speech recognition (IEEE, 2009), pp. 4625–4628
22. J. Makhoul, Linear prediction: A tutorial review. Proceedings of the IEEE **63**(4), 561–580 (1975)
23. H. Morgan, N. Bayya, A. Kohn, P. Hermansky, RASTA-PLP speech analysis. ICSI Technical Report TR-91-969 (1991)
24. C. Kim, R.M. Stern, Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **24**(7), 1315–1329 (2016)
25. D. Elenius, M. Blomberg, in *Proceedings from Fonetik*. Dynamic vocal tract length normalization in speech recognition (2010), pp. 29–34 Citeseer
26. B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Communication **53**(9-10), 1062–1087 (2011)
27. S. Lee, S. Yildirim, A. Kazemzadeh, S. Narayanan, in *Ninth European Conference on Speech Communication and Technology*. An articulatory study of emotional speech production (2005)
28. S. Lee, E. Bresch, J. Adams, A. Kazemzadeh, S. Narayanan, in *Ninth International Conference on Spoken Language Processing*. A study of emotional speech articulation using a fast magnetic resonance imaging technique (2006)
29. W.R. Rodrıguez, O. Saz, A. Miguel, E. Lleida, in *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*. On line vocal tract length estimation for speaker normalization in speech recognition (2010)
30. K. Veselý, A. Ghoshal, L. Burget, D. Povey, in *Interspeech*. Sequence-discriminative training of deep neural networks, vol 2013 (2013), pp. 2345–2349
31. M. Bashirpour, M. Geravanchizadeh, Speech emotion recognition based on power normalized cepstral coefficients in noisy conditions. Iran J Electrical Electronic Eng **12**(3), 197–205 (2016)
32. M. Bashirpour, M. Geravanchizadeh, Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments. EURASIP J Audio Speech Music Process **2018**(9), 1–13 (2018)

33. T.K. Kim, T test as a parametric statistic. Kor J Anesthesiol **68**(6), 540–546 (2015)
34. G. Hinton et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine **29**(6), 82–97 (2012)
35. P. Dighe, A. Asaei, H. Bourlard, On quantifying the quality of acoustic models in hybrid DNN-HMM ASR. Speech Communication **119**, 24–35 (2020)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.