Neural Comput & Applic (2016) 27:2615–2628
DOI 10.1007/s00521-015-2030-9

ORIGINAL ARTICLE

Voice conversion system using salient sub-bands and radial basis function

Jagannath Nirmal[1] · Mukesh Zaveri[2] · Suprava Patnaik[3] · Pramod Kachare[4]

Received: 9 February 2013 / Accepted: 9 August 2015 / Published online: 25 August 2015

Abstract The objective of voice conversion is to replace the speaker-dependent characteristics of the source speaker so that it is perceptually similar to that of the target speaker. The speaker-dependent spectral parameters are characterized using single-scale interpolation techniques such as linear predictive coefficients, formant frequencies, mel cepstrum envelope and line spectral frequencies. These features provide a good approximation of the vocal tract, but produce artifacts at the frame boundaries which result in inaccurate parameter estimation and distortion in re-synthesis of the speech signal. This paper presents a novel approach of voice conversion based on multi-scale wavelet packet transform in the framework of radial basis neural network. The basic idea is to split the signal acoustic space into different salient frequency sub-bands, which are finely tuned to capture the speaker identity, conveyed by the speech signal. Characteristics of different wavelet filters are studied to determine the best filter for the proposed voice conversion system. A relative performance of the proposed algorithm is compared with the state-of-the-art wavelet-based voice morphing using various subjective and objective measures. The results reveal that the proposed algorithm performs better than the conventional wavelet-based voice morphing.

Keywords Discrete wavelet transforms · Dynamic time warping · Wavelet packet transform · Radial basis function · Sub-bands · Voice conversion

✉ Jagannath Nirmal
jhnirmal1975@gmail.com

Mukesh Zaveri
mazaveri@coed.svnit.ac.in

Suprava Patnaik
ssp@eced.svnit.ac.in

Pramod Kachare
pramod_1991@yahoo.com

[1] Department of Electronics Engineering, KJSCE, Mumbai, India

[2] Department of Computer Engineering, SVNIT, Surat, India

[3] Department of Electronics Engineering, SVNIT, Surat, India

[4] Department of Electrical Engineering, VJTI, Mumbai, India

1 Introduction

The aim of the voice conversion is to modify the characteristics of source speaker utterance so that it impersonates the target speaker utterance. Voice conversion has many applications in the areas such as customization of text to speech, speaker dubbing, health care, karaoke, broadcasting and multimedia applications [1–3]. The voice conversion system needs to identify the features relevant to voice individuality and modify them in such a way that the modified speech signal sounds natural and is perceived as if spoken by a target speaker [4]. There are various single-scale speech features which are used to represent vocal tract. They can be classified into three different categories, namely first category of features that belong to acoustic phonetic models such as formant frequencies and formant bandwidth [5]; second category of features derived without considering the speech models such as mel cepstrum envelope [4, 6], cepstrum coefficients and mel-frequency cepstrum coefficients (MFCCs) [7]; and third category of features which uses a parametric approach including linear predictive coefficients (LPC) [8], reflection coefficients [9], log area ratio [8] and line spectral frequencies (LSF) [1, 2, 10–12]. Techniques using LP-related features assume

stationary characteristics of the speech signal within a frame and therefore fail to analyze the local speech variation accurately. Also LPC techniques cannot capture nostril and unvoiced sounds [13]. The MFCC is one of the dominating techniques to capture the speaker-specific features of the speech signal, due to its sub-band-based processing using multi-scale filter bank. However, in the synthesis stage MFCC loses pitch and phase-related information [14, 15].

Various speaker-specific models have been found in the literature, and amongst them, vector quantization (VQ)-based codebook mapping and Gaussian mixture model (GMM) are the most primitive approaches for transformation of vocal tract characteristics [1, 16–19]. In VQ-based technique, the speakers voice signals are clustered and the mapping rule for each cluster is formed using minimum mean square error (MSE). But the main drawback of this technique is hard partitioning, which produces discontinuities in the transition regions, and therefore, it affects the quality and naturalness of converted speech signal [19]. Fuzzy vector quantization [6] and a speaker transformation algorithm using Segmental Codebook (STASC) [2] are proposed to overcome the above limitations. Dynamic frequency warping (DFW) transformation technique is used to improve the quality of converted speech. This DFW technique translates the formants to the new frequencies without modifying the complete spectral shape, which results in poor-quality speech signal [9]. In GMM-based approaches, the quality of converted speech signal is improved by modeling the joint distributions of source and target speech features. In this GMM-based technique, the speakers spectral space is partitioned into overlapping classes and a continuous probabilistic linear transformation function is defined from these partitions for parametric vector representation of envelope [17]. However, the quality and the naturalness of the converted speech signal are found to be inadequate due to reconstruction of speech signal using the large number of parameters which results in over-smoothing problem [13, 20]. To overcome the reconstruction and over-smoothing problem of GMM, different approaches such as speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) [19], harmonic noise model (HNM) [16], phase reconstruction and post-filtering [7] methods are proposed. The over-smoothing and reconstruction problem is partially improved using GMM with weighted frequency warping [21]. The novel speech synthesis technique based on hidden Markov model (HMM) is also proposed for voice conversion. This system generates parameter vector sequences. When a text input is given to a trained HMM [22] set, speech signal reconstruction can be done. Thus, the voice conversion is done by adopting HMMs [23] to the target speaker. However, the quality of the reconstructed speech is limited due to reconstruction and over-smoothing problems similar to that of GMM-based voice conversion. Over-fitting problem of GMM is overcome using partial least squares regression technique [24].

Apart from these, various artificial neural networks (ANN) are proposed to capture the acoustical nonlinearities between source and target speakers [4, 11, 25–28]. The wavelet transform is extensively used for signal analysis and synthesis. Initially, sub-band-based approach is proposed for voice transformation [29]. Wavelet-based approach is used for voice morphing [11] by considering only the low-frequency contents. In this approach, removal of high-frequency contents introduces the muffed effect in synthesized speech signal [30]. An auditory sub-band-based wavelet neural network architecture is proposed for voice conversion [31]. This architecture approximates the human auditory system, which is widely used for speech classification [31]. However, voice conversion requires speaker-specific characteristics to be properly fitted for stimulating transformation model [15]. Most of the speech-related information is uniformly distributed in fundamental frequency and its harmonics (i.e., formants). The first three significant formants are encoded in 200–3 kHz frequency band [32], whereas speaker-specific characteristics are distributed non-uniformly in higher-frequency bands, which are the cause of different articulatory speech organs [13]. The glottis information is encoded between low-frequency band from 100 to 400 Hz, and the piriform fossa information is positioned in medium-frequency band (around 4 kHz). Another speaker-specific constriction is the consonant factors that exist in higher-frequency region (around 7 kHz) [13].

In this paper, we have proposed a wavelet packet filter structure that analyzes the speech signal without considering any underlying knowledge of the human auditory system. A logical way to design proposed system is to derive the speaker-specific characteristics confined in different sub-bands and treat them separately. The salient sub-band-based feature set is derived to capture the speaker-specific characteristics. The wavelet packet transform is combined with the radial basis neural network (RBFNN) to accomplish the nonlinearity between source and target salient sub-bands. The contribution of this paper is to: (1) explore the characteristics of different wavelet filters to determine the best match for the proposed voice conversion system, (2) propose salient multi-scale wavelet packet sub-band-based feature set to modify the acoustic cues of source speaker into target speaker and (3) design RBF-based transformation model to capture the nonlinearity between the source and the target feature sets.

The remainder of this paper is structured as follows: The next section describes the selection of wavelet packet

transform. Section 3 describes salient sub-band selection methodology. The proposed algorithm is explained in Sect. 4. Section 5 enlightens the design of RBF-based voice conversion system. Experimentation results and evaluations are reported in Sect. 6. Conclusions and discussion are given in Sect. 7.

2 Wavelet packet transform

The main motivation of using multi-scale wavelet packet transform (WPT) is its ability to isolate the speaker-specific information from the speech signal to overcome the inefficiency of single-scale features. WPT repetitively divides the wideband input signal into narrowbands by passing it through low-pass and high-pass filters. Equal data rate is maintained in all sub-bands with the use of sampling units at each decomposition level [31].

WPT decomposes the input signal in the series of basis functions called as wavelets, which are denoted as $\Psi_{a,b}(t)$. The variables a and b are scale and translation parameters of the corresponding wavelet. The basis functions $\Psi_{a,b}(t)$ are generated from the mother wavelet $\Psi(t)$ by scaling and translation,

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right) \tag{1}$$

where $\frac{1}{\sqrt{a}}$ is energy normalization in different scales.

In the proposed voice conversion algorithm, we have used WP instead of discrete wavelet transform (DWT) for sub-band decomposition of input speech signal because Heisenbergs uncertainty principle results in a logarithmic frequency resolution. It limits the application of DWT in noisy speech environment and also degrades the speech quality. Unlike WT, WPT decomposes input speech signal not only in low-frequency branches (i.e., approximation coefficients) but also in high-frequency branches (i.e., detailed coefficients) at each level of decomposition. Therefore, WPT with superior frequency localization is used to segment input broadband signal into narrowband signals [33–35].

The factors responsible for the choice of a particular wavelet packet transform are: symmetry, regularity and number of vanishing moments [34]. Symmetry deals with linear-phase finite impulse response (FIR) in the digital filter design for signal reconstruction. Since the quality of converted speech signal after reconstruction is an integral part of the voice conversion system, symmetry becomes prime requirement for the synthesis stage. Regularity also seems to be very important in voice conversion as it deals with smoothness of the transform and has cosmetic influence of smoothing error during the reconstruction. Increase in number of vanishing moments represents more support and insignificant detailed coefficients in higher order. This provides better representation of the signal using approximation coefficients. Wavelet basis with above required characteristics gives us the choice of four wavelet filters, namely Daubechies, symlet, biorthogonal and coiflet [31, 35, 36].

Different speech samples collected from male and female are decomposed up to the fifth level using above-mentioned wavelet families and again re-synthesized. The best wavelet is selected using normalized mean squared error (*NMSE*) criteria [30] between original speech signal $y(i)$ and reconstructed signal $y^*(i)$ with sample length N and frame index i, which is calculated as,

$$\text{NMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(y(i)-y^*(i)\right)^2}{\sum_{i=1}^{N}y(i)^2}} \tag{2}$$

We have calculated NMSE for different wavelet, and it is found that the coiflet5 wavelet produces minimum *NMSE* 1.204 for male and biorthogonal 6.8 produces 1.21 for female speaker, and therefore, for rest of the implementation, coiflet5 and biorthogonal 6.8 have been used.

3 WPT-based salient sub-band feature extraction

Speech signal possesses speech message content and speaker identity. It is essential to identify and incorporate speaker-specific information in the design of the voice conversion system [25]. The speaker-specific characteristics can be isolated from the speech signal by WPT via sub-band decomposition of speech signal. Due to sub-band decomposition of speech, information is localized in different frequency bands. In addition to energy measure, entropy is used to select the salient sub-bands. To obtain the speaker-specific salient sub-bands, the 100 utterances of different speakers are taken from the ARCTIC database, which is sampled at 16 kHz (i.e., 8 kHz bandwidth), and after preprocessing (framing and windowing), each pre-processed frame is decomposed using WPT up to at most the fifth level [34]. Normalized energy and entropy concentration of each sub-band is computed at each approximation and detail level [36, 37]. In general, 90 % of voiced speech energy is concentrated in the first $N/2$ levels in N-level decomposed wavelet transformed sub-bands [38]. The normalized energy of all sub-bands shown in Fig. 1 represents that the lower sub-bands in the range of 0–4 kHz carry most of the speech phonemic discriminative glottis and resonant frequencies of the speech signal. But to preserve the naturalness and speaker-specific information such as piriform fossa, consonant constriction factors and quality of the speech signal, higher-frequency bands need to be considered. For the selection of higher-frequency bands,

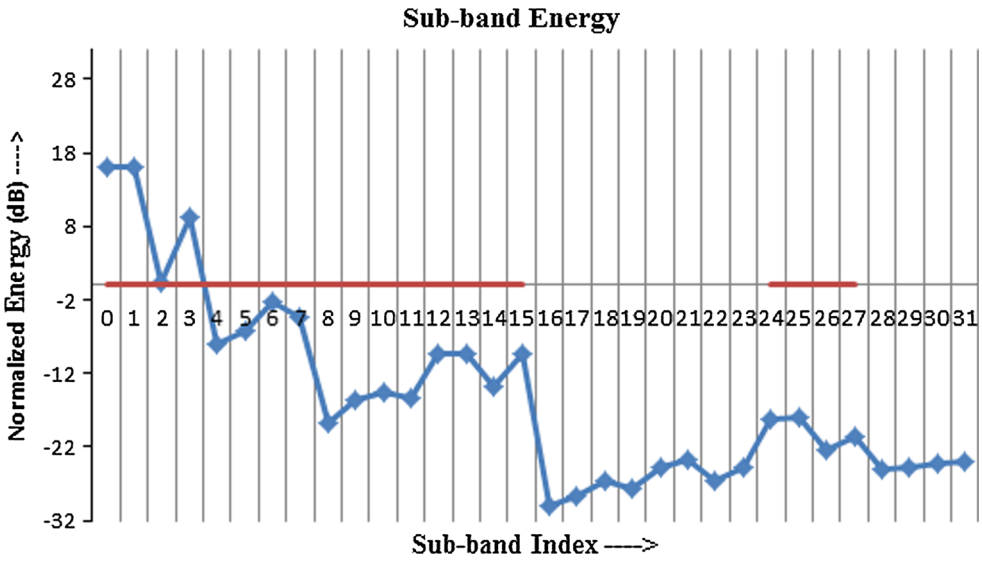

Fig. 1 Average energy content of each sub-band from 100 speech samples

Fig. 2 Average entropy of each sub-band from 100 speech samples

entropy criteria are used. An energy criterion alone is inadequate, as all the high-frequency bands carry low energies. Discrimination between different iso-energetic high-frequency sub-bands can be made by using sub-band entropy calculation shown in Fig. 2. Conferring to Shannons information theory [39], Shannon entropy measures the predictable value of the information contained in a signal. Considering a random variable Y with k conclusions $y_1, \ldots, y_k$, the Shannon entropy $H(Y)$ is defined as,

$$H(Y) = -\sum_{i=1}^{k} p(y_i) \log(p(y_i)) \tag{3}$$

In this equation, $p(y_i)$ is the probability density function for ith conclusion. In the same way, considering histogram of WPT sub-bands for different bin widths [40], the histogram approach uses the idea that the differential entropy can be approximated by producing a histogram of the frequency bins and then finding the discrete entropy of the histogram [41, 42], which is itself a maximum-likelihood estimate of the discretized frequency distribution, where w is the width of the ith bin.

$$H(Y) = -\sum_{i=1}^{k} f(y_i) \log\left(\frac{f(y_i)}{w(y_i)}\right) \tag{4}$$

The Shannon entropy can be calculated for the extracted wavelet packet sub-bands, using Eq. (4). This quantity, in some sense, will evaluate the amount and the rate of information, produced by a process that is represented as a discrete information source. Therefore, sub-bands having higher entropy are selected from high-frequency bands of 6–7 kHz. In lower sub-bands (5.0–5.15), the energy concentration is more than 40 %, and these speech segments

are voiced or combination of voiced and unvoiced [38]. In the medium-frequency bands, the energy concentration is less than that of 40 %, and these speech samples are found as unvoiced part. The other constriction consonant factors are distributed in higher bands [38]. Extreme sub-bands (5.28–5.31) are excluded as they are mostly noise impaired. This reduces the optimal number of sub-bands to 20 (i.e., 5.0–5.15 and 5.24–5.27 as shown in Fig. 3). Energy distribution of salient sub-bands shown in Fig. 1 confirms that 99.76 % energy is confined in these salient sub-bands. Finally, we have designed these salient sub-bands by wavelet packet decomposition of each frame carried up to two levels. This partitions the frequency axis into four bands (0–2, 2–4, 4–6 and 6–8 kHz) each of 2 kHz bandwidth. The frequency bands of 0–2 and 2–4 kHz are further decomposed up to three levels with the bandwidth of 500 Hz. The band in the frequency range of 6–7 kHz is also further decomposed up to two levels with 500 Hz bandwidth each as shown in Fig. 3. The detailed procedure for selection of salient sub-bands is illustrated in Table 1.

The synthesized speech signal is reconstructed from salient sub-bands, and subjective listening test is performed to confirm originality and high quality of the signal.

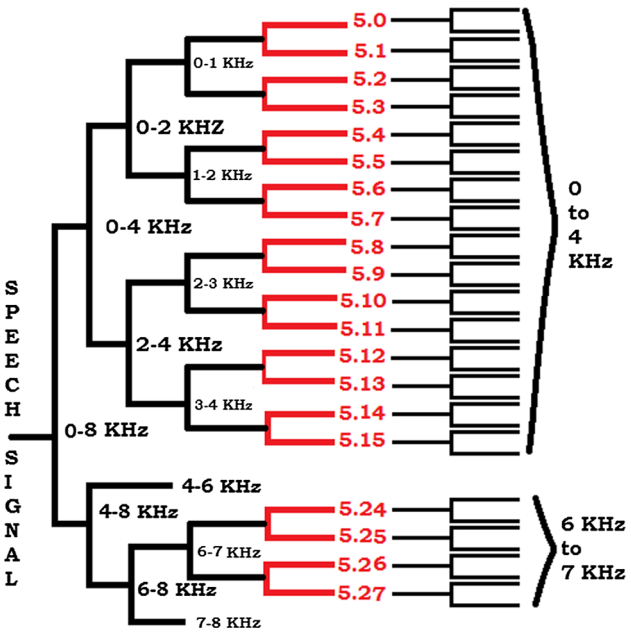

Fig. 3 Proposed wavelet filter bank for selection of salient sub-bands

4 Proposed model

The functional block diagram of the proposed voice conversion algorithm is depicted in Fig. 4. It consists of two phases, namely (1) training phase and (2) testing phase. During the training phase, the beginning and ending silence periods of each phonetically balanced parallel utterance of source and target speakers are removed using voice activity detection (VAD) technique [24]. The residual signal is normalized to have zero mean and unit variance. The training samples of source and target speaker were segmented into frames of 24 m sec (i.e., 400 samples per frame) with 50 % overlap to maintain high quality during reconstruction. Each frame of the source and target frame is decomposed up to fifth level using WPT.

At fifth-level decomposition, a total of 32 sub-bands are calculated, out of which only 20 sub-bands are considered from each source and target speech frame (as discussed in Sect. 3). This procedure is repeated for all utterances of source and target directories. Usually, the length of source and target feature vectors are different so dynamic time

Table 1 Steps to find salient sub-bands

1. Segment source speech signal in frames of 24 msec (i.e., 400 samples/frame) each with 50 % overlapping
2. Decompose each frame up to 5th level in 32 different sub-bands using best match wavelet packet transform
3. Calculate normalized sub-band energy of individual frame of each sub-band [36] as,

$$E_k = 10\log_{10}\left\{\frac{\sum_{i=1}^{M_k}\left[W_k^P x(i)\right]^2}{M_k}\right\}(dB), \qquad k = 1, 2, \ldots, N \tag{5}$$

   where $W_k^P x(i)$ is the wavelet packet transform of signal x at W_k^P node, i is the sub-band frequency index, N is the total number of nodes, M_k is the number of coefficients in the kth sub-band
4. Compute the Shannon entropy [37] of each sub-band as,

$$H_k = -\sum_{i\forall N} p_i * \log(p_i) \tag{6}$$

   where, p_i is the probability that wavelet coefficients (W_k^p) can be located within a bin and N is the number of partitions in histogram of the coefficients space, k is sub-band index
5. Select optimum number of sub-bands based on higher energy, entropy of wavelet coefficients in different sub-bands as shown in Figs. 1, 2
6. Reconstruct the individual frame using selected sub-bands while setting remaining sub-bands coefficients as zero vectors
7. Concatenate reconstructed frames to form a complete speech signal using overlap-add method
8. Compute NMSE and perform listening test to confirm that the selected sub-bands represent original signal efficiently between original and reconstructed speech

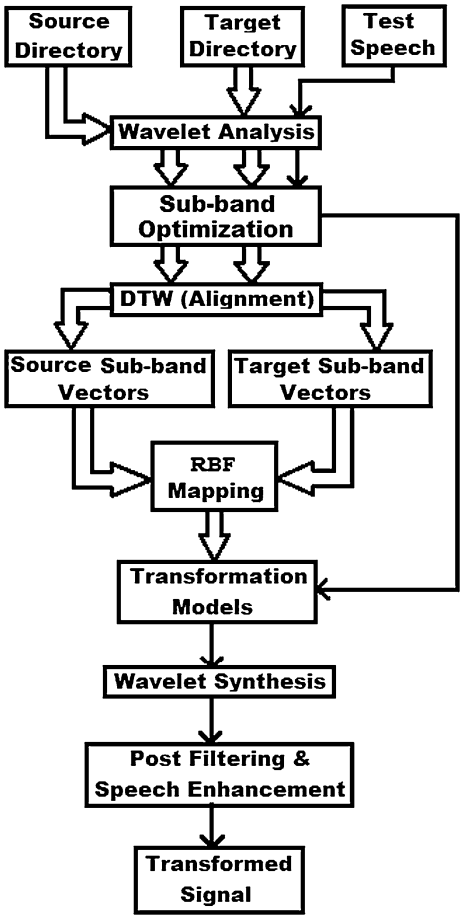

Fig. 4 Proposed architecture for voice conversion

warping is used to align it [16]. After the alignment, source and target feature vectors are normalized and used as training set to develop the RBFN-based mapping function to capture the nonlinear relationship between source and target speaker [37]. RBFN is described in the following section. Several models are explored to decide the best transformation model for proposed voice conversion system. In order to obtain the best transformation model, several RBF models are explored for proposed voice conversion system. The testing phase is followed by training phase.

In testing phase, the parallel utterances of test speaker are preprocessed and split into an optimum number of sub-bands. Feature vector of test speaker is obtained with the procedure similar to that of the training set feature vector. In order to produce transformed sub-band coefficients, the test speaker feature vector is projected through the trained RBFN model. These coefficients are then de-normalized and combined with 12 zero vector sub-bands to reconstruct the frames using inverse wavelet transform. Speech signal reconstruction is accomplished through overlap-add method to retain its original size. Speech enhancement is made with post-filtering blocks. Similar process is repeated for all other test signals. The transformed speech signal contains the characteristics of the target speaker.

5 Radial basis function for mapping

The RBF neural network is a special case of feed forward network, which maps input space nonlinearly to hidden space followed by linear mapping from hidden space to output space. The network represents a map from M_0-dimensional input space to N_0-dimensional output space written as, $S : R^{M_0} \rightarrow R^{N_0}$. When a training dataset of input output pairs $[xk, dk]$, $k = 1, 2 \ldots M_0$, is presented to the RBFNN model, the mapping function F is computed as,

$$F_k(x) = \sum_{j=1}^{N} w_{jk}\phi\left(\left\|x - d_j\right\|\right) \tag{7}$$

where $\|.\|$ is a norm usually Euclidian and computes the distance between applied input x and training data point d_j. Above equation can also be written in matrix form as [26],

$$Fx = W\phi \tag{8}$$

where $\phi(\|x - d_j\|)$, $j = 1, 2 \ldots N$, in which N is the set of arbitrary functions known as radial basis functions. The commonly considered form of ϕ is Gaussian function defined as [26],

$$\phi(x) = \mathrm{e}^{\frac{\|x-\mu\|^2}{2\sigma^2}} \tag{9}$$

RBFNN learning process includes training and generalized phase. The training phase constitutes the optimization of basis function parameters using only input dataset with k-means algorithm in an unsupervised manner. In the second phase, hidden-output neurons weights are optimized in a least square sense by minimizing squared error function,

$$E = \frac{1}{2}\sum_{n}\sum_{k}\left[f_k(x^n) - (d_k)^n\right]^2 \tag{10}$$

where $(d_k)^n$ is desired value for kth output unit when input to the network is x^n. The weight vector is determined as,

$$W = \phi TD \tag{11}$$

where ϕ: matrix of size $(n \times j)$, D: matrix of size $(n \times k)$, ϕ^T: transpose of matrix ϕ.

$$\left(\phi^T\phi\right)W = \phi^T D \tag{12}$$

$$W = \left(\phi^T\phi\right)^{-1}\phi^T D \tag{13}$$

where $\left(\phi^T\phi\right)^{-1}\phi^T$ is pseudo-inverse of matrix ϕ, D is $(d_k)^n$. The weight matrix W can be calculated by linear

inverse matrix technique and used for mapping between the source and target acoustic feature vector. Effective functioning of the RBFNN needs to select optimized kernel parameters which include kernel centers and spread factor. In our work, we have calculated spectral distortion [19] for different kernel spread factors and hidden neurons. We have selected the spread factor of 0.01 with lowest spectral distortion.

6 Experimental results

In order to train the RBF-based mapping functions, the phonetically balance CMU ARCTIC corpus [43] is used. For this experimentation, we have used samples of four speakers, AWB (M1), CLB (F1), SLT (F2) and BDL (M2) from the database. Using these samples developed the different speaker combinations of M1–F1, F2–M2, F1–F2 and M1–M2 for voice conversion. The performance of proposed and baseline techniques is evaluated using different objective and subjective measures.

6.1 Objective evaluation

In this work, various objective measures such as mel cepstral distortion (MCD), performance index (P-LSF), formant deviation, formant distortion and spectrogram are considered.

The MCD is correlated with subjective test results so it is considered for the evaluation. The MCD between the converted speech and target speech is calculated as [4, 24],

$$\mathrm{MCD} = 10\log\left(\sqrt{\sum_{i=1}^{D}\mathrm{mcc}^{ta_i} - \mathrm{mcc}^{tr_i}}\right) \tag{14}$$

where mcc^{ta_i} and mcc^{tr_i} are the ith mel cepstrum coefficients (MCC) of the target and transformed speech, respectively. The zeroth term is not considered in MCD computation as it describes the energy of the frame and it is usually copied from the source.

The performance of voice conversion system experimentally tests for different number of training samples obtained from source and target speakers of male and female, respectively. Figure 6 shows the MCD score for different trained RBF models for M1–F1 are developed. Similarly, the transformation models for M1–M2, F1–F2 and F2–M2 for different numbers of parallel utterances (ranging from 2 to 500) of respective source and target speakers are developed.

Figure 5, shows the MCD obtained for M1–F1 and F2–M2, i.e., inter-gender voice conversion is lesser than the M1–M2 and F1–F2, i.e., intra-gender voice conversion. We also observe from Fig. 5 that the MCD values of RBF network decrease with an increase in the number of training samples.

The performance index (PLSF) is calculated for exploring the requirement of normalized error between the various speaker combinations. The spectral distortion between the target and converted samples, $D_{\mathrm{LSF}}(d(n), d(n))$, and the inter speaker spectral distortion, $D_{\mathrm{LSF}}(d(n), s(n))$, are employed for computing the PLSF measure. Generally, the spectral distortion between speech signals u and v, $D_{\mathrm{LSF}}(u, v)$, is computed as,

$$D_{\mathrm{LSF}}(u,v) = \left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{P}\sum_{j=1}^{P}\left(\mathrm{LSF}_u^{i,j} - \mathrm{LSF}_u^{i,j}\right)^2}\right] \tag{15}$$

where N denotes the number of frames, P denotes LSF order and $\mathrm{LSF}_u^{i,j}$ is the jth LSF coefficients in the frame i. The P_{LSF} measure is defined as,

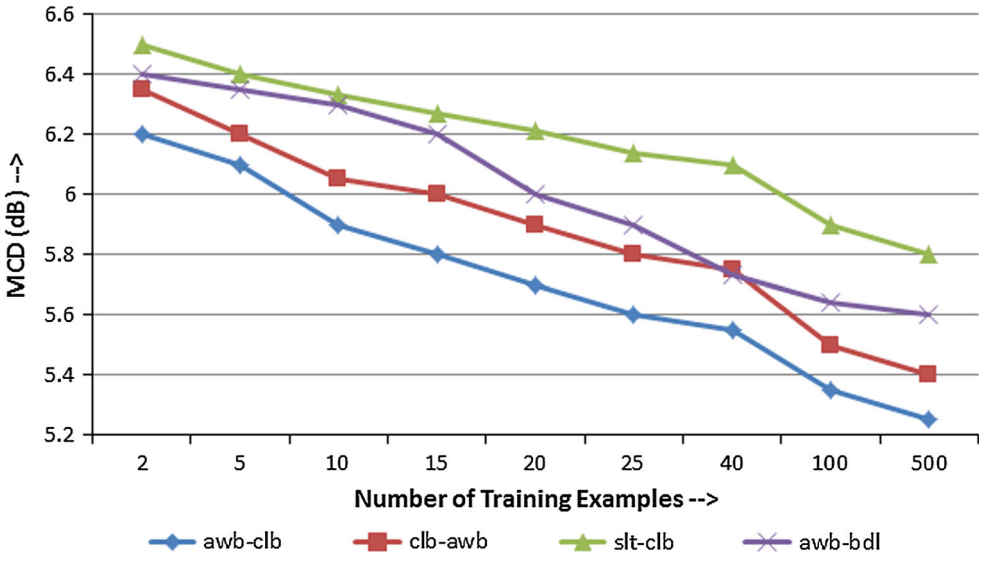

Fig. 5 Performance of RBF model for different source and target transforming pairs

$$P_{\mathrm{LSF}} = \left[1 - \frac{D_{\mathrm{LSF}}(d(n), \hat{d}(n))}{D_{\mathrm{LSF}}(d(n), s(n))}\right] \tag{16}$$

The performance index $P_{\mathrm{LSF}} = 1$ specifies that the transformed speech signal is indistinguishable to the desired one, whereas $P_{\mathrm{LSF}} = 0$ identifies that the transformed speech signal is not at all related to the desired one.

The performance index operates on the input–output parameters of the transformation function, and it directly describes the performance of the transformation model. In the computation of this index, four different converted samples for each speaker combination of M1–F1, F2–M2, M1–M2 and F1–F2 are considered. Table 2 shows that the performance of M1–F1 in proposed voice conversion is more effective than the other conversion combinations. From Table 2, it is clear that the performance of the proposed salient sub-band algorithm is more effective than the baseline wavelet-based voice morphing using RBF.

Along with MCD and P-LSF, different objective measures such as deviation (D_k), root mean square error (μ_{RMSE}) and correlation coefficients $(\gamma_{x,y})$ are also calculated for same speaker combinations. Deviation is defined as the percentage variation in the desired (x_k) and predicted (y_k) formant frequencies obtained from the speech frames. It corresponds to the percentage of test frames within a specified deviation. Deviation (D_k) is computed as,

$$D_k = \frac{|x_k - y_k|}{x_k} \times 100 \tag{17}$$

The root mean square error is calculated as percentage of average of desired formant values attained from the speech segments.

$$\mu_{\mathrm{RMSE}} = \frac{\sqrt{\frac{\sum_k |x_k - y_k|^2}{N}}}{\bar{x}} * 100 \tag{18}$$

$$\sigma = \sqrt{\sum_k d_k^2}, d_k = e_k - \mu, e_k = x_k - y_k, \quad \mu = \frac{\sum_k |x_k - y_k|}{N} \tag{19}$$

where the error e_k is the difference between the actual and predicted formant values, N is the number of observed formant values of speech frames and the parameter d_k is the error in the deviation. The correlation coefficient $\gamma_{(x,y)}$ is the parameter which is to be determined from the covariance $\mathrm{COV}(X, Y)$ between the target (x) and the predicted (y) formant values and the standard deviations σ_X, σ_Y of the target and the predicted formant values, respectively. The parameters $\gamma_{(x,y)}$ and $\mathrm{COV}(X, Y)$ are calculated using Eq. (20),

$$\gamma_{x,y} = \frac{\mathrm{COV(X, Y)}}{\sigma_X \sigma_Y}, \mathrm{COV}(X, Y) = \frac{\sum_k |(x_k - \bar{x})(y_k - \bar{y})|}{N} \tag{20}$$

The objective measures, namely deviation (D_i), root mean square error (RMSE) and correlation coefficients $(\gamma_{(x,y)})$ of M1–F1, F2–M2, F1–F2 and M1–M2, are obtained for state-of-the-art wavelet-based algorithm and given in Table 3. Similarly, Table 4 shows the measures obtained for proposed voice conversion algorithm. From the tables, it can be observed that the μ_{RMSE} between the desired and the predicted acoustic space parameters for proposed model is less than the baseline model. However, every time RMSE does not give strong information about the spectral distortion. Consequently, scatter plots and spectral distortion are employed additionally as objective evaluation measures.

For evaluation of both the salient sub-band-based RBF mapping function and wavelet-based voice morphing, various samples of intra-gender and inter-gender voice conversion are considered. For each speech frame, the desired speakers LSFs are predicted, and from that the corresponding LPCs and formant frequencies are derived. All these objective measures are tabulated for each of the speaker combinations for M1–F1, F2–M2, F2–F1 and M1–M2. First column of Tables 3 and 4 shows the formant frequencies from f1 to f4. Column 3–9 indicate the percentage of speech frames predicting the formant frequencies within specified deviation, and column 10 and 11 specify the RMSE and correlation coefficients, respectively (Fig. 6).

The prediction performance of the optimized RBF models for converting the salient sub-bands and baseline wavelet-based approach is demonstrated using scatter plots. For development of these scatter plots, different utterances are selected randomly from the test samples. The actual and predicted formants frequencies are derived

Table 2 Comparative performance indices of different speaker combinations for proposed and baseline wavelet-based approach

	Sample 1		Sample 2		Sample 3		Sample 4	
	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline
M1–F1	0.7815	0.5072	0.7626	0.4731	0.7706	0.5547	0.7504	0.5118
F2–M2	0.7685	0.4631	0.7922	0.4582	0.7603	0.4764	0.7531	0.4904
M1–M2	0.7249	0.4121	0.7306	0.4408	0.7193	0.4492	0.7403	0.4892
F1–F2	0.7156	0.4024	0.7236	0.4335	0.7375	0.4867	0.7309	0.5480

Table 3 Performance of baseline wavelet-based voice morphing for predicting formant frequencies within a specified percentage of deviation

Transformation model	Formant frequencies	% Predicted frame within deviation								
		2 (%)	5 (%)	10 (%)	15 (%)	20 (%)	25 (%)	50 (%)	μ_{RMSE}	$\gamma_{x,y}$
M1–F1	f1	56	76	82	87	88	89	92	4.36	0.74
	f2	40	61	77	79	83	85	90	3.63	0.78
	f3	22	45	61	66	70	73	89	3.25	0.71
	f4	7	13	23	40	52	65	93	3.05	0.67
F2–M2	f1	51	65	71	77	79	82	91	3.92	0.65
	f2	44	64	72	77	82	84	92	3.47	0.57
	f3	29	48	59	65	70	73	88	3.31	0.22
	f4	6	19	39	53	63	74	94	2.91	0.26
F1–F2	f1	53	69	80	84	86	90	91	4.28	0.69
	f2	38	60	74	81	84	85	89	3.36	0.74
	f3	27	39	58	67	72	73	86	3.51	0.73
	f4	9	16	27	40	51	66	82	3.59	0.71
M1–M2	f1	40	55	66	74	78	80	90	3.98	0.67
	f2	44	46	62	67	72	74	91	3.17	0.6
	f3	27	46	57	63	68	74	82	3.12	0.42
	f4	8	20	38	51	73	80	88	3.97	0.36

Table 4 Performance of proposed salient sub-band-based voice conversion for predicting formant frequencies within a specified percentage of deviation

Transformation model	Formant frequencies	% Predicted frame within deviation								
		2 (%)	5 (%)	10 (%)	15 (%)	20 (%)	25 (%)	50 (%)	μ_{RMSE}	$\gamma_{x,y}$
M1–F1	f1	42	76	86	92	93	94	94	3.26	0.86
	f2	24	62	81	86	90	92	93	3.39	0.82
	f3	58	80	88	91	92	93	98	2.29	0.77
	f4	55	71	82	87	88	91	99	2.34	0.76
F2–M2	f1	30	48	65	73	77	80	87	3.51	0.88
	f2	54	69	75	82	86	87	95	3.74	0.68
	f3	72	82	85	89	91	94	96	3.05	0.71
	f4	51	66	81	86	90	94	100	2.47	0.74
F1–F2	f1	38	73	85	89	95	95	97	4.51	0.58
	f2	41	72	79	82	83	86	90	3.74	0.68
	f3	79	82	85	89	90	93	95	3.05	0.71
	f4	56	68	74	76	79	82	96	2.47	0.74
M1–M2	f1	15	30	50	56	56	60	72	4.51	0.58
	f2	18	34	48	53	59	60	67	3.74	0.68
	f3	23	47	58	64	68	74	90	3.05	0.71
	f4	38	51	61	71	74	86	100	2.47	0.74

from the chosen speech frames jointly and used for the development of these scatter plots. Figures 7 and 8 show the formant frequencies for different speaker combinations, measuring the vocal tract prediction performance for proposed algorithm.

The transformed formant patterns for a specific frame of the target and transformed speech signal are obtained for all speaker combinations using proposed and baseline algorithm. Figure 8 depicts that the pattern of the corresponding transformed signals produced using proposed is closely following the particular target signal, whereas the figure also shows that the predicted formant pattern of baseline approach is closely following the target pattern only for lower formants.

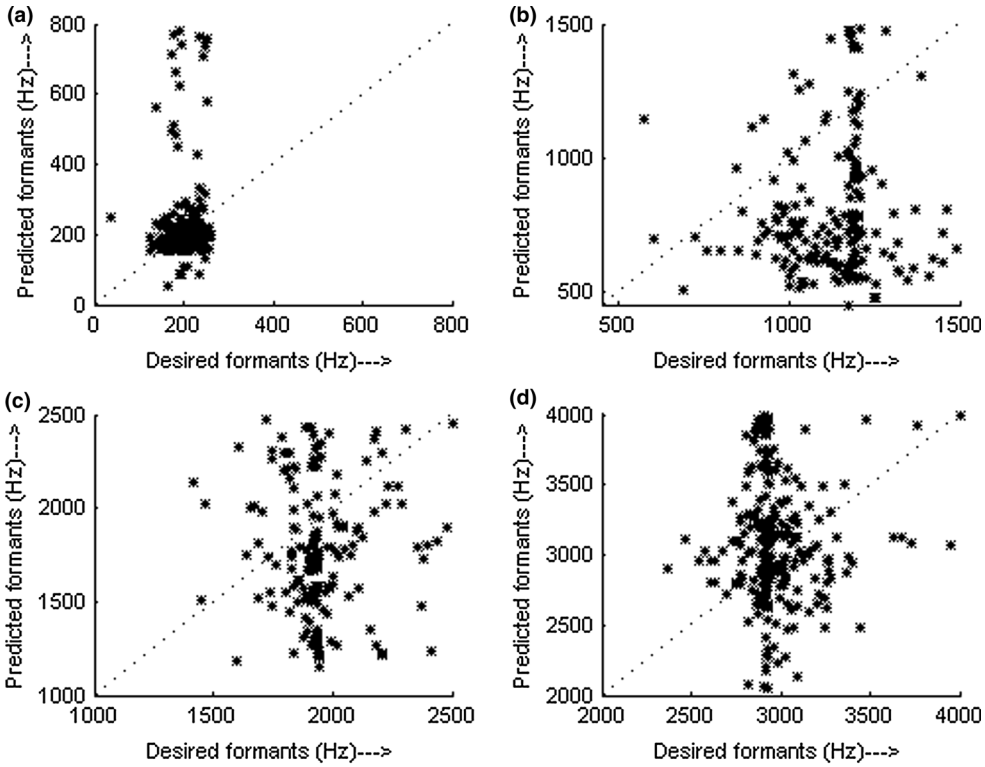

Fig. 6 Desired and predicted values of the formant frequencies of M1–F1 for *a* first, *b* second, *c* third and *d* fourth formants

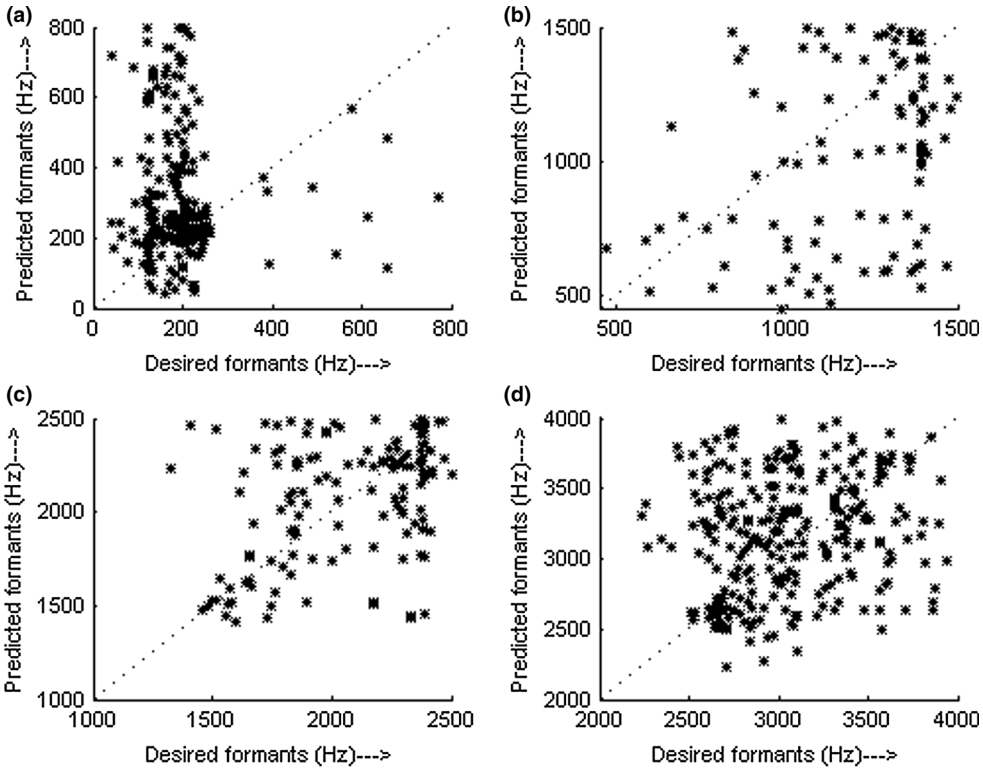

Fig. 7 Desired and predicted values of the formant frequencies of F2–M2 for *a* first, *b* second, *c* third and *d* fourth formants

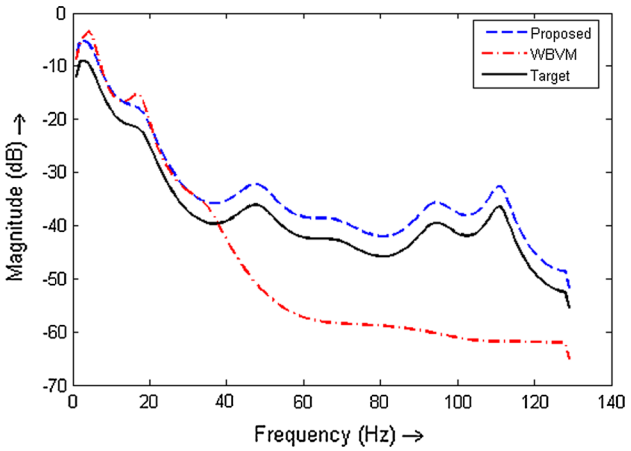

Fig. 8 Comparing spectral envelope for M1–F1 voice conversion

Figure 9 shows the spectrograms of the (a) target speech signal and transformed speech signal of (b) wavelet-based morphing and (c) salient sub-band-based voice conversion. It is clear from the figure that the formant structure of the desired speech signal is almost similar to that of converted speech signal of the proposed algorithm than the baseline algorithm.

6.2 Subjective evaluation

The basic goal of voice conversion system is to modify the source speaker speech so that it mimics the target speaker speech. Therefore, the closeness between the transformed and desired speech signals is evaluated using different subjective listening tests. For inter-gender and intra-gender conversion, different source and target parallel utterances are extracted from the source and target directories and different mapping functions were developed for 2–500 samples. For each one, different utterances are reconstructed from their associated trained functions. The subjective listening tests such as ABX and mean opinion score (MOS) are used to assess the closeness of speech identity and quality between synthesized speech and desired speech, respectively. For these evaluations, we have developed transformation models from 40 parallel utterances. The synthesized speech and their corresponding utterances of target directories presented to the 13 student listeners to judge their comparative performance with corresponding source and target. The student listeners have given their opinion in the scale of 1–5. A speaker individuality test, ABX comprised of A: Source, B: Target and X: Transformed speech signal is also conducted using the same set of utterances. In the ABX test, the listeners are asked to judge which of A or B sounds was closer to X in terms of speaker individuality. Higher the value of ABX, the more the nearness of the transformed speech to the desired utterance. The ABX score 5 of a synthesized speech indicates the exact target speech, whereas score 1 indicates exact source speech. These ratings represent the

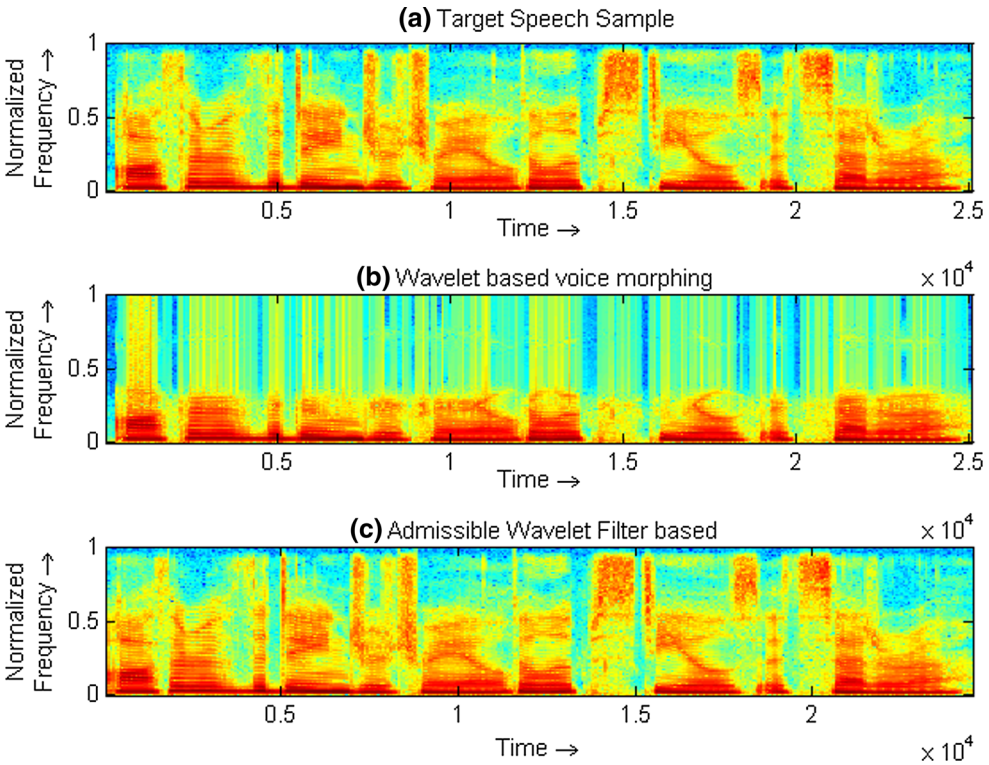

Fig. 9 Spectrogram comparison of M1–F1 voice conversion

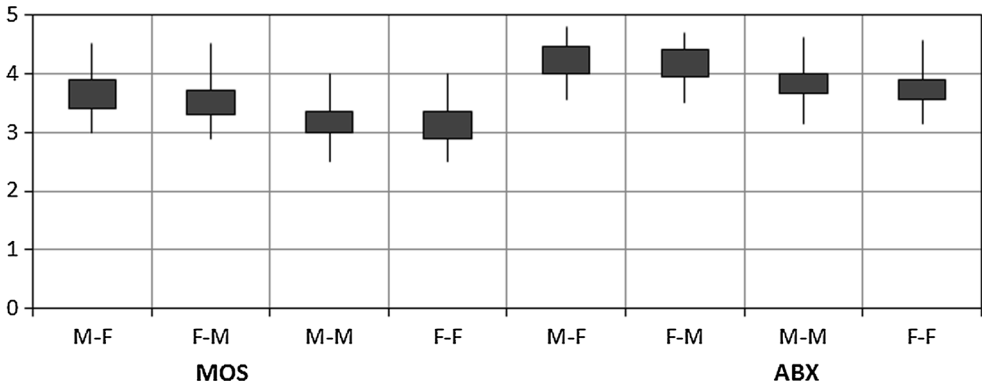

Fig. 10 Result of subjective analysis for similarity and quality in stock plot representation

closeness between source and target on a scale of 1 to 5 as shown in Fig. 10. To assess the speech quality and naturalness of converting a speech signal and transformed speech signal, MOS (i.e., Preference) is conducted, and here listeners were asked to judge the speech quality and naturalness in the rank of 1–5. The MOS score 5 of a converted speech represents high-quality natural utterance, whereas score 1 indicates highly distorted speech signal. The obtained MOS represents the effectiveness of mapping function for inter-gender and intra-gender conversion. The same listeners have given their opinion also shown in Fig. 10. In conclusion, we have compared our subjective analysis with that of the state-of-the-art algorithm [22] and inferred that the perceptual results of the proposed algorithm are superior for inter-gender voice conversion.

In inter-gender (male to female or female to male) conversion, the MOS is more as compared to intra-gender conversion. This MOS variation is clearly reflected with respect to their gender, and the difference in the length of the vocal tract and intonation pattern of inter-gender speaker is large.

7 Conclusion

In this paper, wavelet packet sub-band-based RBF framework is studied for transforming source speaker acoustics into target speaker acoustics. Initially, available wavelet filters above needed constraints are analyzed to select the suitable mother wavelet. Further, 20 finely tuned sub-bands are selected to capture voice individuality, naturalness and quality of speech signals and verified under energy as well as entropy maximization criteria. The RBF-based neural network is established to generalize the relationship between source and target feature vectors. The permutation of source and target speakers helps in generating various transformation models. Multiple objective and subjective measures are employed to justify the improved performance of the proposed over the state-of-the-art voice morphing technique.

The performance of the proposed approach verified the significance of combining the high-frequency information with low-frequency information to use it effectively for voice conversion. Hence, the muffed effect at the output of the state-of-the-art voice morphing technique can be alleviated. The results also reveal that the conversion for source and target speakers of dissimilar genders (inter-gender) performs slightly better while maintaining high speech quality. The optimization of sub-bands in the proposed algorithm reduces the computational complexity and accelerates the network convergence. System performance can further be improved by using phonetically aligned or syllable level aligned database during training phase.

References

1. Kain A, Macon MW (2001) Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction. In: Proceedings of IEEE international conference acoustics speech signal processing, vol 2, pp. 813–816
2. Arslan LM (1999) Speaker transformation algorithm using segmental code books (STASC). Speech Commun 28:211–226
3. Lee K (2007) Statistical approach for voice personality transformation. IEEE Trans Audio Speech Lang Process 15:641–651
4. Rabiner L, Juang BH (1993) Fundamentals of speech recognition. Prentice Hall of India, New Delhi
5. Furui S (1986) Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Commun 5(2):183–197
6. Rao KS (2010) Voice conversion by mapping the speaker-specific features using pitch synchronous approach. Comput Speech Lang Process 24(3):474–494
7. Drioli C (2001) Radial basis function networks for conversion of sound speech spectra. EURASIP J Appl Signal Process 1:36–40
8. Strang G, Nguyen T (1997) Wavelets and filter banks. Wellesley Cambridge Press, Wellesley
9. Valbret H, Moulines E, Tubach JP (1992) Voice transformation using PSOLA technique. Speech Commun 1:145–148
10. Chadha AN, Nirmal JH, Kachare P (2014) A Comparative performance of various speech analysis-synthesis techniques. Int J Signal Process Syst 2(1):17
11. Deshpande Mangesh S, Holambe Raghunath S (2010) Speaker identification using admissible wavelet packet based decomposition. World Acad Sci Eng Technol 37:736–739

12. Nirmal JH, Zaveri M, Patnaik S, Kachare P (2014) Voice conversion using general egression neural network. Appl Soft Comput 24:1–12
13. Narendranath M, Murthy HA, Rajendran S, Yegnanarayana B (1995) Transformation of formants for voice conversion using artificial neural networks. Speech Commun 16(2):207–216
14. Stylianou Y (1996) Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification, Ph.D. dissertation, cole Nationale Superieure Des Tlcommunications. Paris, France
15. Stylianou Y, Capp O, Moulines E (1998) Continuous probabilistic transform for voice conversion. In: Proceedings IEEE international conference acoustics, speech, signal process. vol 6, pp. 131–142
16. Nirmal JH, Zaveri M, Patnaik S, Kachare P (2014) Complex cepstrum based voice conversion using radial basis function neural network. In: ISRN signal processing, vol 2014. Hindawi Publishing Corporation, Article ID 357048
17. Kominek J, Black AW (2004) The CMU ARCTIC speech databases. In: Proceedings 5th ISCA speech synthesis workshop (SSW5), Pittsburgh, PA, pp. 223–224
18. Guidoa Rodrigo C, Vieiraa Lucimar Sasso, Juniora Sylvio Barbon (2007) A neural wavelet architectures for voice conversion. Sci Direct Neurocomput 71:174–180
19. Kuwabura H, Sagisaka Y (1995) Acoustic characteristics of speaker individuality: control and conversion. Speech Commun 16:165–173
20. Laskara RH, Chakrabartyb D, Talukdara FA, Sreenivasa Raoc K, Banerjeea K (2012) Comparing ANN and GMM in a voice conversion framework. Appl Soft Comput 12:3332–3342
21. Stylianou Y, Cappe Y, Moulines E (1998) Continuous probabilistic transform for voice conversion. IEEE Trans Speech Audio Process 6:131142
22. Desai S, Black AW, Yegnanarayana B, Prahallad K (2010) Spectral mapping using artificial neural networks for voice conversion. IEEE Trans Audio Speech Lang Process 18(5):954–964
23. Childers DG, Yegnanarayana B, Wu K (1985) Voice conversion: factor responsible for quality. In: Proceedings of IEEE ICASSP, pp. 530–533
24. Kuwabura H, Sagisaka Y (1995) Acoustic characteristics of speaker individuality: control and conversion. Speech Commun 16:165–173
25. Narendranath M, Murthy HA, Rajendran S, Yegnanarayana B (1995) Transformation of formants for voice conversion using artificial neural networks. Speech Commun 16(2):207–216
26. Nirmal JH, Zaveri M, Patnaik S, Kachare P (2013) A novel voice conversion approach using admissible wavelet packet decomposition. EURASIP J Audio Speech Music Process 2013:28
27. Helander E, Virtanen T, Jani N, Gabbouj M (2010) Voice conversion using partial least squares regression. IEEE Trans Audio Speech Lang Process 18(5):912–921
28. Desai S, Black AW, Yegnanarayana B, Prahallad K (2010) Spectral mapping using artificial neural networks for voice conversion. IEEE Trans Audio Speech Lang Process 18(5):954–964
29. Masuko T, Tokuda K, Kobayashi T, Imai S (1996) Speech synthesis using HMMS with dynamic features. In: Proceedings IEEE international conference acoustics, speech, signal processing, pp. 389–392
30. Orphanidou C, Moroz IM, Roberts SJ (2004) Wavelet-based voice morphing. WSEAS J Syst 10(3):3297–3302
31. Guidoa Rodrigo C, Vieiraa Lucimar Sasso, Juniora Sylvio Barbon (2007) A neural wavelet architectures for voice conversion. Sci Direct Neurocomput 71:174–180
32. Nirmal JH, Patnaik SS, Zaveri MA (2012) Voice transformation using radial basis function. In: Third international conference on recent trends in information. Telecommunication and computing ITC 2012, Springer, Berlin, pp. 271–276
33. Furui S (1986) Research on individuality features in speech waves and automatic speaker recognition techniques. Speech Commun 5(2):183–197
34. Strang G, Nguyen T (1997) Wavelets and filter banks. Wellesley Cambridge Press, Wellesley
35. Rao KS (2010) Voice conversion by mapping the speaker-specific features using pitch synchronous approach. Comput Speech Lang Process 24(3):474–494
36. Deshpande Mangesh S, Holambe Raghunath S (2010) Speaker identification using admissible wavelet packet based decomposition. World Acad Sci Eng Technol 37:736–739
37. Nirmal JH, Zaveri M, Patnaik S, Kachare P (2013) A novel voice conversion approach using admissible wavelet packet decomposition. EURASIP J Audio Speech Music Process 2013:28
38. Xugang Lu, Dang Jianwu (2008) An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Commun 50:312–322
39. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423
40. Shannon CE (2001) A mathematical theory of communication. ACM SIGMOBILE Mob Comput Commun Rev 5(1):3–55
41. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E 69(6):066138
42. Xugang Lu, Dang Jianwu (2008) An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Commun 50:312–322
43. Reza Fazlollah M (1961, 1994) An introduction to information theory. Dover Publications Inc., New York