EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

A novel voice conversion approach using admissible wavelet packet decomposition

Jagannath H Nirmal[1]*, Mukesh A Zaveri[2], Suprava Patnaik[1] and Pramod H Kachare[3]

Abstract

The framework of voice conversion system is expected to emphasize both the static and dynamic characteristics of the speech signal. The conventional approaches like Mel frequency cepstrum coefficients and linear predictive coefficients focus on spectral features limited to lower frequency bands. This paper presents a novel wavelet packet filter bank approach to identify non-uniformly distributed dynamic characteristics of the speaker. Contribution of this paper is threefold. First, in the feature extraction stage, dyadic wavelet packet tree structure is optimized to involve less computation while preserving the speaker-specific features. Second, in the feature representation step, magnitude and phase attributes are treated separately to rule out on the fact that raw time-frequency traits are highly correlated but carry intelligent speech information. Finally, the RBF mapping function is established to transform the speaker-specific features from the source to the target speakers. The results obtained by the proposed filter bank-based voice conversion system are compared to the baseline multiscale voice morphing results by using subjective and objective measures. Evaluation results reveal that the proposed method outperforms by incorporating the speaker-specific dynamic characteristics and phase information of the speech signal.

Keywords: Admissible wavelet packet; Dynamic time warping; Radial basis function; Speaker-specific features; Wavelet-based filter bank

1 Introduction

The voice conversion (VC) system aims to apply various modifications to the source speaker's voice so that the converted signal sounds like a particular target speaker's voice [1,2]. The VC system is comprised of two phases: training and transformation. The training phase includes feature extraction and incorporates features to formulate an appropriate mapping function. Subsequently, the source speaker characteristics are transformed to that of target speaker using mapping function developed in the training phase [3]. In order to extract the speaker-specific features, several speech feature representations have been developed in the literature, such as Formant Frequencies (FF) [1,4], Linear Predictive Coefficients (LPC) [1,5] and Line Spectral Frequency (LSF) [6-8], Mel Frequency Cepstrum Coefficient (MFCC) [9], Mel Generated Cepstrum (MGC) [10], and spectral lines [11]. The LPC features can provide a good approximation model for the vocal tract characteristics, but it neglects few significant details of the individual speaker, like the nasal cavity, unvoiced sound, and other side branches related to non-linguistic information [12]. For the enhancement of the speech quality, a STRAIGHT approach has been proposed [13]. However, it needs enormous computation and therefore is inappropriate for real-time applications. The methods based on the vocal tract model have been developed using MFCC features considering the nonlinear mechanism of the human auditory system [14]. Most of the above approaches provide a good approximation to the source-filter model. However, these methods ignore fine temporal details during the extraction of formant coefficients and the excitation signal [12,15]. This gives muffled effect in synthesized target speech.

Further improvements in the synthesized speech quality have been reported in various multiscale approaches [16-18]. To our knowledge, initially, the wavelet-based sub-band model proposed by Turk and Arslan produced promising results [16]. Following the ideas of sub-band-based approach, the multiscale approach has been proposed for voice morphing [17]. Afterwards, the auditory

*Correspondence: jhnirmal1975@gmail.com
[1] Department of Electronics Engineering, S.V. National Institute of Technology, Surat 395007, India
Full list of author information is available at the end of the article

sub-band-based wavelet neural network architecture has been proposed [18], which is widely application for speech classification [12]. However, VC needs to model the speech and speaker-specific characteristics of the speech for developing transformation model [4,19]. The features representing speaker identity are distributed non-uniformly in different frequency regions.

This paper presents the wavelet filter structure for extracting the speaker-specific features without considering any underlying knowledge of the human auditory system. This filter bank is analysed using admissible wavelet transform as it gives freedom to decompose the low- and high-frequency bands. The contribution of this paper are as follows: (1) the first is the use of the admissible wavelet packet transform based filter bank to extract the speaker-specific information of the speech signal, (2) the second is to reduce the computational complexity of the proposed features using Discrete Cosine Transform (DCT), (3) the third is to incorporate phase of the DCT coefficients to emphasize that phase equally contributes to the synthesized speech signal naturality as the magnitude.

Radial Basis Function is explored to establish the non-linear mapping rules for modifying the source speaker features to that of the target speaker. The RBF model is used as a mapping model because of its fast training procedure and good generalization properties. Finally, the performance of the proposed filter bank-based VC model is compared with the state-of-the-art multiscale voice morphing using RBF analysis. This is done using various objective measures, such as performance index (P_{LSF}) [1], formant deviation [7,20], and spectral distortion [20]. The commonly used subjective measures such as Mean Opinion Score (MOS) and ABX are used to verify the quality and similarity of the converted speech signal [21].

The rest of the paper is structured as follows: The optimal filter bank is explained in Section 2. The new VC system based on optimal filter bank along with the state of the art multiscale method is explained in Section 3. Thereafter, Section 4 briefs the theoretical aspects of RBF-based transformation model. The database and performance measures for comparison of quality and similarity of the synthesized speech are mentioned in Section 5. Finally, the conclusions are derived in Section 6.

2 Optimal filter bank

The voice individuality caused by different articulatory speech organs is distributed non-uniformly in some invariant parts of the vocal tract, such as the nasal cavity, piriform fossa, and laryngeal tube [12]. The information of the glottis is encoded in the low-frequency region from 100 to 400 Hz, and the piriform fossa is positioned in the medium frequency band from 4 to 5 kHz. The information of consonant factor exists in a higher frequency region, i.e., 7 to 8 kHz [12,14]. The first three formants are encoded in the lower and middle frequency regions from 200 Hz to 3 kHz.

The VC system needs to realize the transformation model considering the speaker-specific features [22]. The traditional auditory filter bank is not suitable to capture the speaker individuality of the speech signal [12,23]. Therefore, the frequency resolution in different bands is restructured considering the non-uniform distribution of the speaker-specific information in these bands. Additional details about wavelet analysis can be found in [17,18,24].

For the design of filter bank, the ARCTIC database is used. The input speech signal sampled at 16 kHz is pre-processed in various stages, such as pre-emphasis, framing, and windowing. The 8-kHz bandwidth speech frame is decomposed up to four levels by wavelet packet decomposition. This partitions the frequency axis into 16 bands each of 500-Hz band width. The different frequency bands with the speaker-specific features are further decomposed to get finer resolution than the Mel filter bank [24,25]. The lower frequency range 0 to 1 kHz captures the fundamental frequency which has maximum energy with most speaker-specific information. Therefore, the lower two bands 0 to 0.5 kHz and 0.5 to 1 kHz is decomposed up to the seventh level. It splits the band of 0 to 1 kHz into 16 sub-bands 62.5 Hz each, which is finer than corresponding bandwidth of Mel filter bank [24]. In addition, the frequency band of 1 to 3 kHz contains the speaker-specific information about the first and second harmonics of the fundamental frequency [23]. This frequency band carries less speaker-specific information compared to previous lower sub-bands. Therefore, the band of 1 to 2 kHz is decomposed up to six levels and 2- to 3-kHz band is decomposed up to five levels. This gives 12 sub-bands with finer frequency resolution than the Mel sub-bands. The frequency band of 4 to 5 kHz related to the invariant part of the vocal tract gives information about the piriform fossa. It holds features suitable for speaker identity. However, the resolution of this frequency range is coarser in Mel filter bank [12]. Therefore, this band is further decomposed up to fifth level. The frequency bands 3 to 4 kHz and 5 to 8 kHz do not require further decomposition as these bands already have a fine frequency resolution than the corresponding bands of Mel filter bank. The significant band decomposition is continued till the substantial energy of the corresponding bands is achieved.

The selection of the wavelet basis is done using root-mean-square error (RMSE) measure [17,18]. In reference with the above discussion, the experiments are carried out. The final filter structure shown in Figure 1 gives best results. It consists of 40 different sub-bands. The quality and naturalness in the VC system may be improved by capturing speaker-specific features in the high-frequency region.

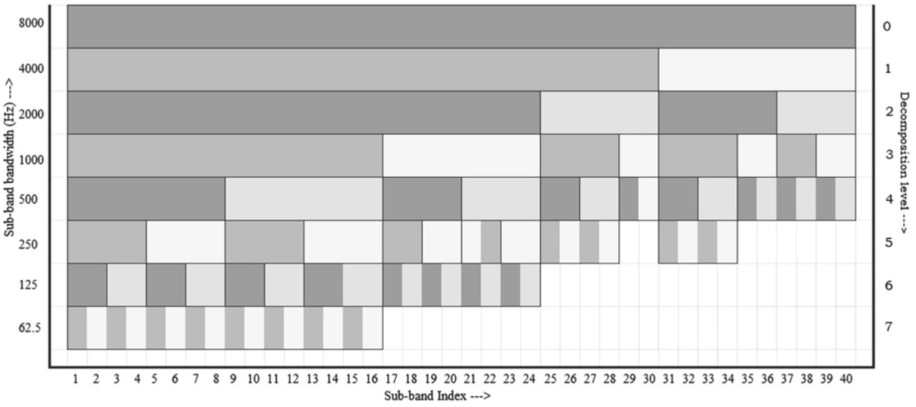

Figure 1 **Proposed filter bank realized using admissible wavelet packet decomposition.**

3 Voice conversion framework

In this section, the design of a new VC algorithm using the proposed filter bank is explained. In order to compare the performance of the proposed VC system with the state-of-the-art multiscale-based voice morphing using RBF analysis [17] is considered.

3.1 Proposed filter bank-based VC

The proposed VC system depicted in Figure 2, consists of two phases: training and transformation. During the training phase, the normalized utterances of the source speaker are segmented into frames of 32 ms each frame consisting of 480 samples. Thereafter, the proposed filter bank is applied to each of these frames. Then, log and DCT transformation of the filter coefficients is carried out to reduce the computational complexity. The feature vector is formed considering the phase along with the magnitude of DCT coefficients [26]. The similar set of procedures is used to obtain the feature vectors from the target speaker. However, it is unlikely that synchronized feature vectors would be obtained even if the source and target speaker utter the same sentence. Therefore, feature vectors of source speaker are time aligned with that of the target speaker to train the mapping model. The alignment is carried out using dynamic time warping technique [9]. The aligned magnitude and phase feature vectors of source and target speakers are used to train the separate RBF-based transformation model to establish conversion rules.

In the transformation stage, the test utterances of source speaker are pre-processed in the similar way as the training stage to get the separate feature vectors for magnitude and phase information of filtered coefficients. Then, the transformed coefficients are obtained by projecting coefficients through the separate trained models. Afterwards, inverse mathematical operations such as Inverse Discrete Cosine Transform (IDCT) and antilog are applied to the transformed coefficients analogous to operations performed in the training phase. The time domain speech frames are computed in the inverse filtering stage

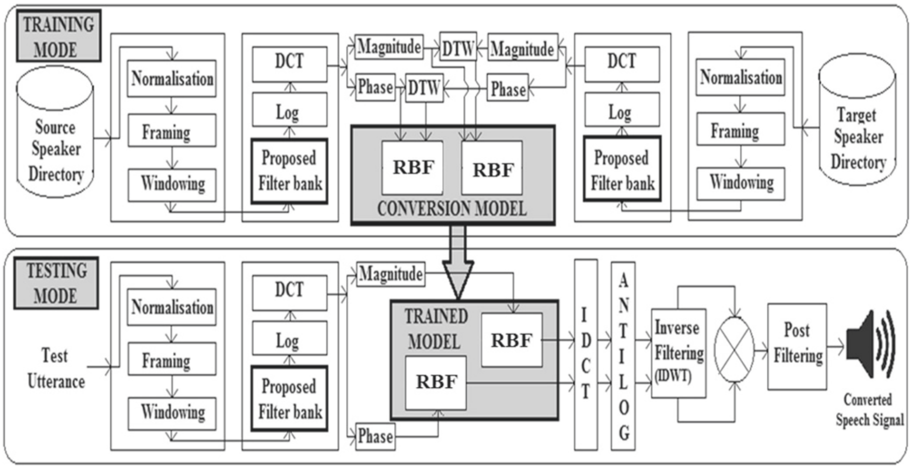

Figure 2 **Block diagram of filter bank-based VC with detailed (a) training and (b) testing mode.**

and combined using overlap-add technique. The use of post filtering followed by inverse operations ensures the good quality of the converted speech signal.

3.2 Baseline multiscale voice morphing using RBF analysis

As discussed earlier, the performance of the proposed algorithm is compared with a state of the art multiscale voice morphing [17]. The pre-processing operations of this method are similar to proposed voice conversion method. The dyadic wavelet filter bank applied to the source and target speech frames partitions each of the frames into different frequency bands. Wavelet basis functions, coiflet5, and bi-orthogonal6.8 are chosen for male-to-female and female-to-male conversion, respectively. The wavelet basis with minimum reconstruction error is chosen. It is important to note that the wavelet coefficients at the highest sub-band are set to zero in this filter bank. The networks are trained using frames of normalized wavelet coefficients of the remaining four levels. The network with minimum error on the validation data is chosen for each level and mapping function at that corresponding level is established [17]. The transformation phase employs the RBF-based mapping rules developed in the training stage to obtain the morphed features of the target speaker [27]. Then, inverse mathematical operations analogous to the training stage are used to reconstruct the target speaker's speech signal.

4 Radial basis function-based mapping

Radial basis function-based transformation model is explored to capture the nonlinear dynamics of the acoustical cues between source and desired target speakers. The baseline method performs spectral conversion using RBF-based transformation model and a similar approach is used in this paper for transforming speaker-specific features [17,28]. 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, $S : R_0^M \rightarrow R_0^N$. The training dataset includes input output pairs $[x_k, d_k]; k = 1, 2 \ldots M_0$. When the M_0 dimensional input x is applied to the RBF model, the mapping function F is computed as [27]

$$F_k(x) = \sum_{j=1}^{m} w_{jk} \Phi(||x - d_j||), \tag{1}$$

where $||.||$ is a norm usually Euclidean, computes the distance between applied input x and training data point d_j. The above equation can also be written in matrix form as

$$F(x) = W\Phi, \tag{2}$$

where $\Phi(||x - d_j||), j = 1, 2 \ldots m$ is the set of m arbitrary functions known as Radial Basis Functions. The σ is the spread factor of the basis function. The commonly considered form of Φ is Gaussian function defined as,

$$\Phi(x) = e^{\frac{||x-\mu||^2}{2\sigma^2}}. \tag{3}$$

The radial basis function network model learning process includes training and generalization phase. Training of the network is carried out using the input dataset alone. The optimized basis function is used in the training phase which is usually obtained using k-means algorithm in an unsupervised manner. In the second phase, the weights in the hidden to output layer are optimized in a least square sense by minimizing squared error function,

$$E = \frac{1}{2} \sum_{m} \sum_{k} [f_k(x^n) - (d_k)^n]^2, \tag{4}$$

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

$$W = \Phi^T D, \tag{5}$$

where Φ: matrix of size (n × j), D: matrix of size (n × k), Φ^T: transpose of matrix Φ,

$$\Phi^T \Phi) W = \Phi^T D \tag{6}$$

$$W = (\Phi^T \Phi)^{-1} \Phi^T D, \tag{7}$$

where $(\Phi^T \Phi)^{-1} \Phi^T$ denotes the pseudo inverse of matrix Φ and D denotes the target matrix for d_k^n. The weight matrix W can be calculated by linear inverse matrix technique and used for mapping between the source and target feature vectors. The exact interpolation of RBF is acquainted with two serious problems namely, poor performance for noisy data and increased computational complexity [28]. These problems can be addressed by modifying two RBF parameters. First, one is the spread factor calculated as

$$\sigma_j = 2 \times \text{avg}||x - \mu_j||. \tag{8}$$

The optimized spread factor confirms that the individual RBFs are neither wide nor narrow. The second is bias unit. A bias unit is introduced into the linear sum of activations at the desired output layer to compensate difference between the mean over the data set of the basis function activations and the corresponding mean of the targets. Hence, we obtain the RBF network with the mapping function $F_k(x)$ computed as

$$F_k(x) = \sum_{j}^{m} w_{jk} \Phi(||x - d_j||). \tag{9}$$

Table 1 Performance index of proposed method and baseline method for different synthesized speech samples

Type of conversion	Performance index							
	Converted sample 1		Converted sample 2		Converted sample 3		Converted sample 4	
	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline	Proposed	Baseline
M1-F1	0.5286	0.4070	0.5372	0.4371	0.5983	0.5547	0.5211	0.5118
F1-M2	0.4486	0.3341	0.4682	0.4582	0.996	0.4764	0.3906	0.3104
F1-F2	0.4249	0.4508	0.4337	0.4408	0.4435	0.4492	0.4348	0.2892
M1-M2	0.5452	0.5003	0.6264	0.5735	0.6924	0.6867	0.5658	0.5480

The separate conversion models are used for mapping the magnitude and phase feature vectors of the source speaker to that of the target speaker. The optimum networks obtained through the training are used to predict the transformed parameters of the target speaker from the source speaker.

5 Experimental results

The training set includes phonetically balanced English utterances of seven professional narrators. The utterances in this database are sampled at 16 kHz. The corpus includes sentences of JMK (Canadian male), BDL (American male), AWB (Scottish male), RMS (American male), KSP (Indian male), CLB (American female), and SLT (American female) [29].

The utterances of two male speakers, AWB (M1) and BDL (M2), and two female speakers, CLB (F1) and SLT (F2), are employed in the analysis. The transformation models are developed for four different speaker combinations: M1-F1, F2-M2, F1-F2, and M1-M2. The minimum 40 parallel utterances are required to form a VC model [9]. Our training set includes 50 parallel utterances obtained from each of the speaker pairs and a separate set of 25 utterances of each source speaker are used to evaluate the system. In order to evaluate the VC system the objective measures, such as performance index, spectral distortion and formant deviations are considered. The end user of the VC system is human so the objective evaluations are confirmed with subjective evaluations. The subjective evaluations involve rating the system performance in terms of similarity and quality of the converted and target speech. Usually, ABX and MOS tests are employed to evaluate similarity and quality, respectively. The performance index (P_{LSF}) is computed for investigating the requirement of normalized error for different pairs. The spectral distortion between desired and transformed utterances, $D_{\mathrm{LSF}}(d(n), \hat{d}(n))$ and the inter speaker spectral distortion, $D_{\mathrm{LSF}}(d(n), s(n))$ are used for computing the P_{LSF} measure. In general, the speaker spectral distortion between signals u and v, $D_{\mathrm{LSF}}(u, v)$ is defined as

$$D_{\mathrm{LSF}}(u, v) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{1}{P} \sum_{j=1}^{P} (\mathrm{LSF}_u^{i,j} - \mathrm{LSF}_v^{i,j})^2}, \tag{10}$$

where N represents the number of frames, P refers to a LSF order, and $\mathrm{LSF}_u^{i,j}$ is the jth LSF component in the frame i. The performance index is given by

$$P_{\mathrm{LSF}} = 1 - \frac{D_{\mathrm{LSF}}(d(n), \hat{d}(n))}{D_{\mathrm{LSF}}(d(n), s(n))}. \tag{11}$$

The performance index $P_{\mathrm{LSF}} = 1$ indicates that the converted signal is identical to the desired one, whereas

Table 2 Performance of baseline method for predicting formant frequencies within a specified percentage of deviation

Transformation model	Formant frequencies	% Predicted frame within deviation							
		2%	10%	15%	20%	25%	50%	μ_{RMSE}	$\gamma(x, y)$
M1-F1	f1	56	82	87	88	89	92	4.36	0.74
	f2	40	77	79	83	85	90	3.63	0.78
	f3	22	61	66	70	73	89	3.25	0.71
	f4	7	23	40	52	65	93	3.05	0.67
F2-M2	f1	51	71	77	79	82	91	3.92	0.65
	f2	44	72	77	82	84	92	3.37	0.57
	f3	29	59	65	70	73	88	3.31	022
	f4	6	39	53	63	74	94	2.91	0.26

Table 3 Performance of proposed method for predicting formant frequencies within a specified percentage of deviation

Transformation model	Formant frequencies	% Predicted frame within deviation							
		2%	10%	15%	20%	25%	50%	μ_{RMSE}	$\gamma_{(x,y)}$
M1-F1	f1	60	87	88	89	90	94	4.46	0.78
	f2	51	83	88	90	91	95	3.39	0.86
	f3	57	90	91	92	94	98	2.78	0.84
	f4	66	89	93	95	96	100	2.41	0.86
F2-M2	f1	35	65	71	76	80	89	3.36	0.71
	f2	44	88	92	93	94	97	2.71	0.86
	f3	58	89	91	94	95	99	2.40	0.74
	f4	81	90	94	96	97	100	2.10	0.77

$P_{\text{LSF}} = 0$ specifies that the converted signal is not at all similar to the desired one.

It can be seen in the Table 1 that both the proposed VC method and baseline method shows performance index differences for M1-F1, F2-M2, M1-M2, and F1-F2 pairs. The results specify that the performance of the proposed system is significantly better than that of the baseline method.

The other performance measures, such as formant deviation (D_k), root-mean-square error (RMSE), and correlation coefficients ($\sigma_{x,y}$) are used to analyse our system. Deviation parameter is defined as, the percentage variation in the actual (x_k) and predicted (y_k) formant frequencies, derived from the corresponding speech frames. It represents the percentage of test frames that lie within a specified deviation (D_k) and is calculated as

$$D_k = \frac{|x_k - y_k|}{x_k} \times 100. \tag{12}$$

For a given transformed and target signals, root-mean-square error is calculated in terms of percentage of

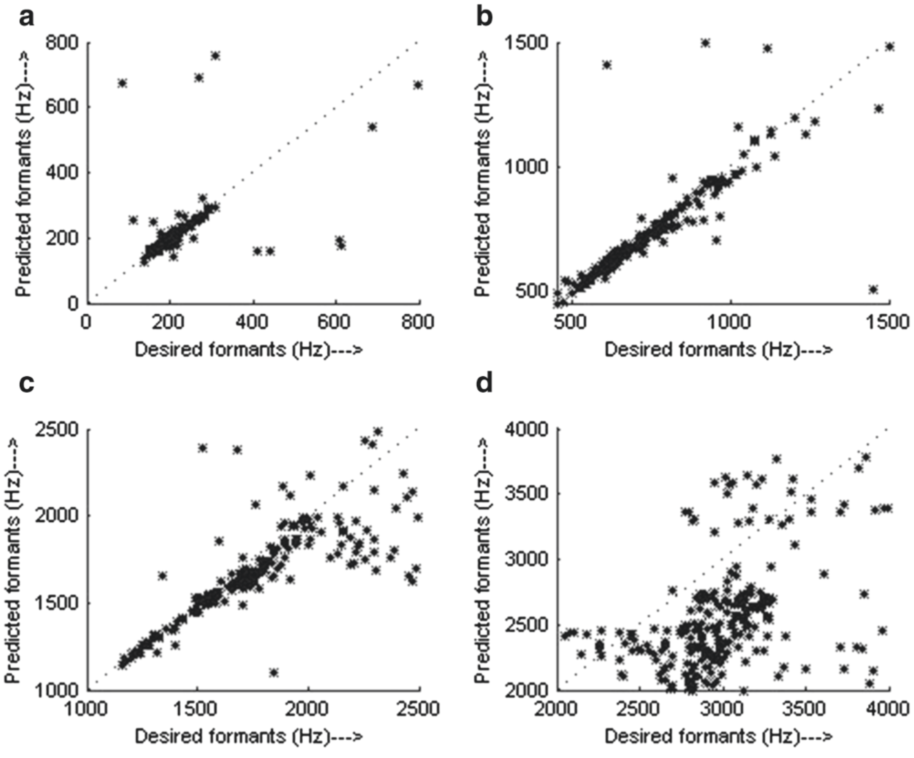

Figure 3 Desired and predicted formant frequencies for M1 to F1 using baseline method. (a) First, **(b)** second, **(c)** third, and **(d)** fourth formants.

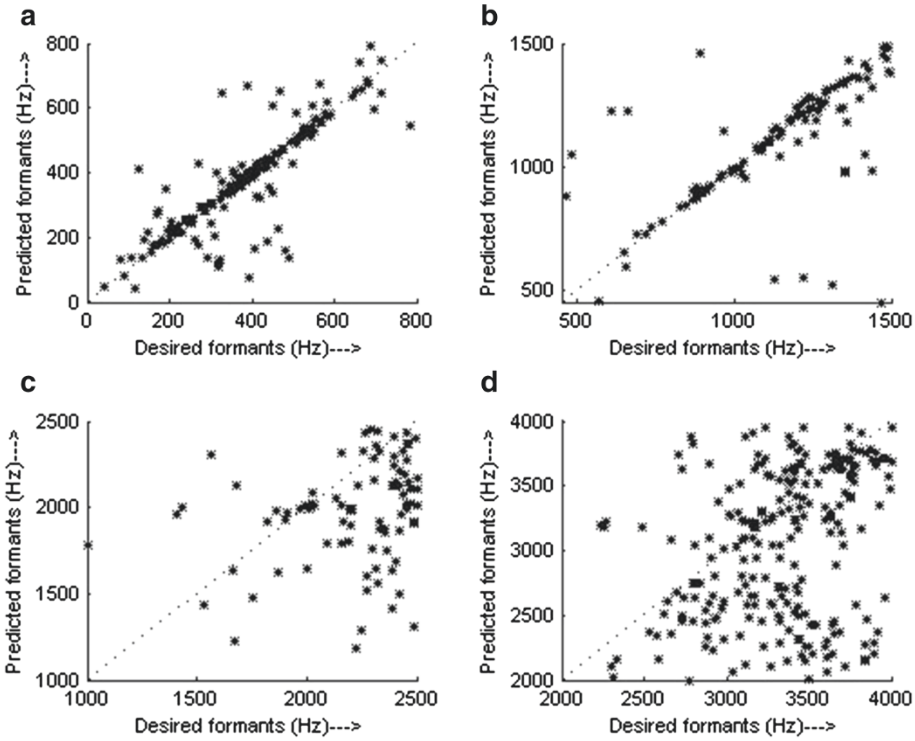

Figure 4 Desired and predicted formant frequencies for F2 to M2 using baseline method. (a) First, **(b)** second, **(c)** third, and **(d)** fourth formants.

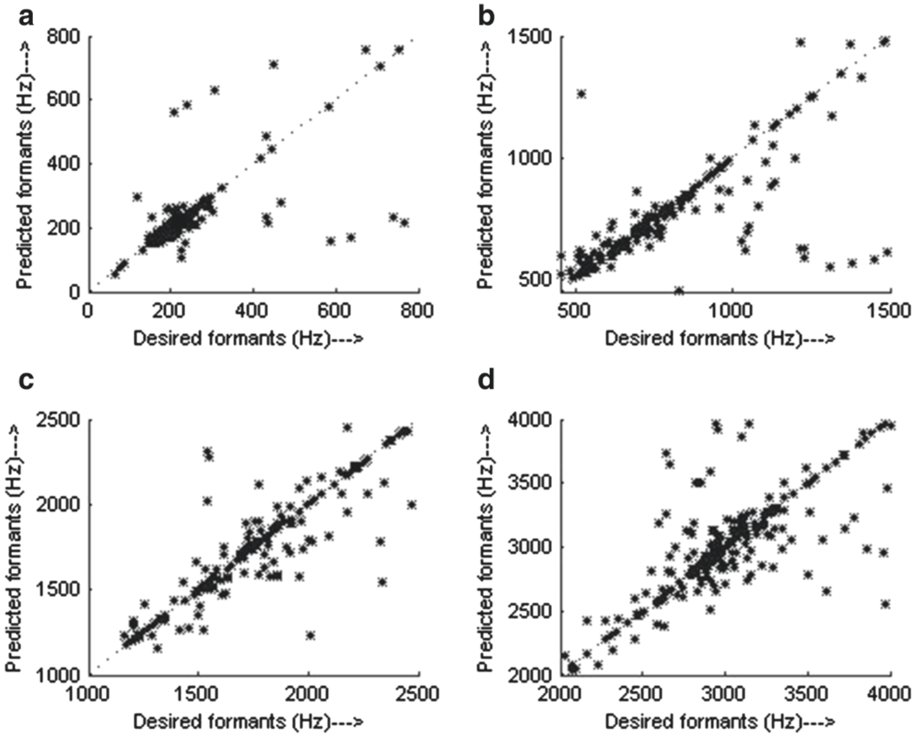

Figure 5 Desired and predicted formant frequencies for M1 to F1 using proposed method. (a) First, **(b)** second, **(c)** third, and **(d)** fourth formants.

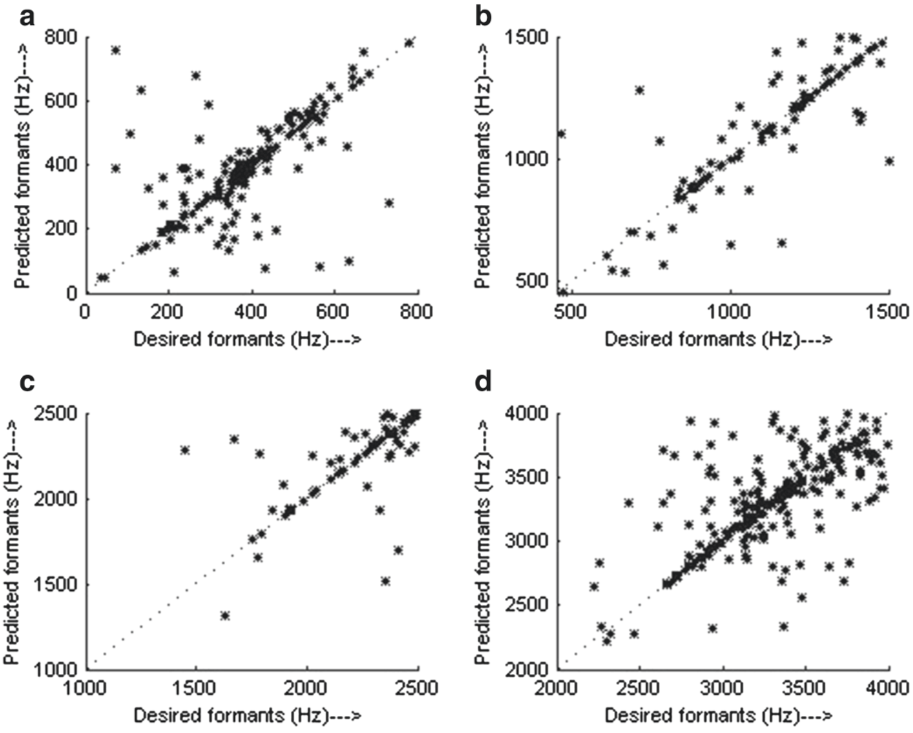

Figure 6 Desired and predicted formant frequencies for F2 to M2 using proposed method. (a) First, **(b)** second, **(c)** third, and **(d)** fourth formants.

average desired formant values obtained from the speech segments. It is computed as

$$\mu_{\text{RMSE}} = \frac{\sqrt{\sum_k |x_k - y_k|^2}}{\bar{x}} \times 100, \tag{13}$$

where $\sigma = \sqrt{\sum_k d_k^2}$

$$d_k = e_k - \mu, e_k = x_k - y_k, \mu = \frac{\sum_k |x_k - y_k|}{N}. \tag{14}$$

The error e_k is the difference between the actual and predicted formant values. N is the number of observed formant frequencies of speech frames. The parameter d_k is the deviation error. The correlation coefficient $\gamma_(x, y)$ is the parameter to be computed in terms of covariance $\text{COV}(X, Y)$ 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 $\text{COV}(X, Y)$ are calculated as

$$\gamma_{x,y} = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y} \tag{15}$$

$$\text{COV}(X, Y) = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{N}. \tag{16}$$

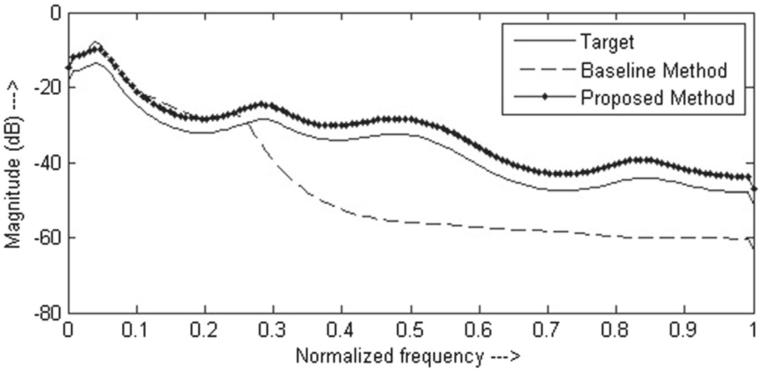

Figure 7 Target and transformed spectral envelopes of the desired speaker using proposed method and baseline method.

Table 4 Score used in speech quality (MOS) and speaker identity (ABX)

Score	MOS (speech quality)	ABX (speaker identity)
1	Bad (imperfect to perceive)	Totally different
2	Poor (almost impossible to perceive)	Certainly not
3	Fair (sound perception is not perfect)	Possibly different
4	Very good (cell phone quality)	More or less the same
5	Outstanding (perfect to perceive)	Totally same

Table 2 shows the different objective measures (D_K), RMSE, and ($\gamma_{x,y}$) for different speaker combinations (M1-F1, F2-M2) using baseline method. Table 3 shows the similar measures for proposed method.

One can observe that the RMSE values between the desired and the predicted acoustic space parameters for proposed model are less than that of the baseline model. However, every time RMSE does not give strong information about the spectral distortion. Consequently, scatter plot and spectral distortion are employed additionally as objective measures.

The baseline method scatter plots for first-, second-, third-, and fourth-order formant frequencies using M1-F1 and F2-M2 speaker pairs are shown in Figures 3 and 4, respectively. Similar analysis is done for proposed method as shown in Figures 5 and 6. The clusters obtained using proposed model are more compact and diagonally oriented compared to the baseline model. It is observed that the higher predicted formants are more closely oriented toward the desired formants for proposed filter bank-based method than that of the baseline method. Also, the diagonal orientation of the clusters demonstrates the good prediction ability of both the methods, as perfect prediction means all the data points in scatter plot are diagonally oriented in right side. The compact clusters obtained for proposed method imply its ability to capture the formant structure of desired speaker.

Figure 7 shows the desired and predicted spectral envelopes for proposed and baseline method. It can be seen in the figure that the spectral envelopes obtained for proposed method follow the same shape and have peaks and valleys at same frequencies confirming the similarity between them. On the other hand, the spectral envelopes for baseline method have different shapes.

Table 5 Subjective analysis for quality (MOS) and identity (ABX)

	Proposed algorithm		Baseline algorithm	
	MOS	ABX	MOS	ABX
M1-F1	4.58	4.88	4.17	4.39
F2-M2	4.59	4.83	4.03	4.23
M1-M2	4.47	4.77	4.24	4.38
F1-F2	4.50	4.79	4.12	4.32

As mentioned earlier, the proposed and the baseline methods are also evaluated in terms of subjective tests: MOS and ABX. Mean opinion score is a quality evaluation test for the synthesized speech and ABX is the test for similarity between converted and target speech signal. The tests related to quality and similarity are carried out using 25 synthesized speech utterances obtained from four different speaker pairs and the corresponding target utterances. In the first part, the listeners are asked to judge the quality of synthesized speech signal using MOS in the scale of 1 to 5 as shown in Table 4. The MOS results shown in Table 5 indicates that the conversion is more in proposed method than baseline method.

In the next part of the evaluation, the ABX similarity test (A: Source, B: Target, X: Transformed speech signal) is carried out without considering the speech quality. The listeners are asked to grade the speaker identity on the five-point scale. The listeners are asked to give ratings in the scale of 1 to 5 to decide whether the output X matches with A or B as shown in Table 4. The higher value of ABX suggests that mapping functions which are developed with proposed and the baseline method can convert the identity of one speaker to the other with acceptable level. Table 5 shows that the listeners have given better rating to the proposed method than that of the baseline method in term of both MOS and ABX test.

6 Conclusion

In this article, a new feature extraction approach based on admissible wavelet packet transform has been proposed. The earlier feature extraction methods focused only on the low-frequency bands without considering the features in the high-frequency bands which are equally important for speaker identity. The proposed method mainly emphasizes the speech signal frequency regions which are important for speaker identity. The features obtained from the proposed filter bank are modified using RBF-based conversion models. Different objective and subjective measures used in our work justifies the performance of proposed and baseline model. The proposed method gives considerably improved results than the baseline method in terms of both the quality and identity of the speaker. The performance of the proposed system proved the significance of combining the information from the high-frequency bands with low-frequency bands to use it effectively for voice conversion.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] Department of Electronics Engineering, S.V. National Institute of Technology, Surat 395007, India. [2] Department of Computer Engineering, S.V. National

Institute of Technology, Surat 395007, India. [3]Department of Electronics Engineering, Veermata Jeejabai Institute of Technology, Mumbai 400031, India.

Received: 26 June 2013 Accepted: 18 November 2013
Published: 10 December 2013

References

1. K-S Lee, Statistical approach for voice personality transformation. Audio, Speech, Lang. Process., IEEE Trans. **15**(2), 641–651 (2007)
2. H Ye, S Young, High quality voice morphing, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, vol. 1, (Montreal, 17–21 May 2004), pp. I-9–12
3. LM Arslan, Speaker transformation algorithm using segmental code books (stasc). Speech Commun. **28**, 211–226 (1999)
4. H Kuwabara, Y Sagisaka, Acoustics characteristics of speaker individuality: control and conversion. Speech Commun. **16**, 165–173 (1995)
5. M Abe, S Nakamura, K Shikano, H Kuwabara, vol. 1, Voice conversion through vector quantization, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP-88)*, (New York, NY, 11–14 April 1988), pp. 655–658
6. A Kain, MW Macon, Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, vol. 2, (Salt Lake City, UT, 7–11 May 2001), pp. 813–816
7. KS Rao, Voice conversion by mapping the speaker-specific features using pitch synchronous approach. Comput. Speech & Lang. **24**(3), 474–494 (2010)
8. O Turk, LM Arslan, Robust processing techniques for voice conversion. Comput. Speech & Lang. **20**(4), 441–467 (2006)
9. S Desai, AW Black, B Yegnanarayana, K Prahallad, Spectral mapping using artificial neural networks for voice conversion. Audio, Speech, and Lang. Process., IEEE Trans. **18**(5), 954–964 (2010)
10. E Helander, T Virtanen, J Nurminen, M Gabbouj, Voice conversion using partial least squares regression. Audio, Speech, Lang. Process., IEEE Trans. **18**(5), 912–921 (2010)
11. D Sundermann, H Hoge, A Bonafonte, H Ney, AW Black, Residual prediction based on unit selection, in *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, (San Juan, 27 Nov 2005), pp. 369–374
12. X Lu, J Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Commun. **50**(4), 312–322 (2008)
13. H Kawahara, I Masuda-Katsuse, de Cheveigné A, Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extractionpossible role of a repetitive structure in sounds. Speech Commun. **27**(3), 187–207 (1999)
14. S Hayakawa, F Itakura, Text-dependent speaker recognition using the information in the higher frequency band, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-94)*, vol.1 (Adelaide, 19–22 Apr 1994), pp. I/137–140
15. S Imai, vol. 8, Cepstral analysis synthesis on the mel frequency scale, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '83*, (Boston, MA, 14–16 April 1983), pp. 93–96
16. O Turk, LM Arslan, Subband based voice conversion, in *International Conference on Spoken Language Processing*, (Denver, CO, 16–20 Sept 2002), pp. 289–292
17. C Orphanidou, IM Moroz, SJ Roberts, Multiscale voice morphing using radial basis function analysis, in *Algorithms for Approximation* (Springer, Berlin Heidelberg, 2007), pp. 61–69
18. RC Guido, L Sasso Vieira, S Barbon Júnior, FL Sanchez, C Dias Maciel, E Silva Fonseca, J Carlos Pereira, A neural-wavelet architecture for voice conversion. Neurocomputing. **71**(1), 174–180 (2007)
19. S Furui, Research of individuality features in speech waves and automatic speaker recognition techniques. Speech Commun. **5**(2), 183–197 (1986)
20. R Laskar, D Chakrabarty, F Talukdar, KS Rao, K Banerjee, Comparing ann and gmm in a voice conversion framework. Appl. Soft Comput. **12**(11), 3332–3342 (2012)
21. J Nurminen, V Popa, J Tian, Y Tang, I Kiss, A parametric approach for voice conversion, in *International (TC-STAR) Workshop on Speech-to-Speech Translation, Audio and Visual Communications,Nokia Research Center*, (Barcelona, Spain, June 2006), pp. 225–229
22. M Narendranath, HA Murthy, S Rajendran, B Yegnanarayana, Transformation of formants for voice conversion using artificial neural networks. Speech Commun. **16**(2), 207–216 (1995)
23. E Ormanci, UH Nikbay, O Turk, LM Arslan, Subjective assessment of frequency bands for perception of speaker identity, in *Proceedings of the ICSLP 2002,INTERSPEECH*, (Denver, CO, 16–20 September 2002), pp. 2581–2584
24. O Farooq, S Datta, Mel filter-like admissible wavelet packet structure for speech recognition. Signal Processing Letters, IEEE **8**(7), 196–198 (2001)
25. S-Y Lung, Wavelet feature selection based neural networks with application to the text independent speaker identification. Pattern Recognit. **39**(8), 1518–1521 (2006)
26. LD Alsteris, KK Paliwal, Short-time phase spectrum in speech processing: a review and some experimental results. Digit. Signal Process. **17**(3), 578–616 (2007)
27. T Watanabe, T Murakami, M Namba, T Hoya, Y Ishida, Transformation of spectral envelope for voice conversion based on radial basis function networks, in *Seventh International Conference on Spoken Language Processing, INTERSPEECH, ISCA(2002),* (Denver, CO, 16–20 September 2002)
28. N Iwahashi, Y Sagisaka, Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. Speech Commun. **16**(2), 139–151 (1995)
29. J Kominek, AW Black, The CMU ARCTIC Speech Databases, in *SSW5-2004*, (Pittsburgh, PA, 14–16 June 2004), pp. 223–224

doi:10.1186/1687-4722-2013-28
Cite this article as: Nirmal *et al.*: A novel voice conversion approach using admissible wavelet packet decomposition. *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:28.