Hindawi Publishing Corporation
ISRN Signal Processing
Volume 2014, Article ID 357048, 13 pages
http://dx.doi.org/10.1155/2014/357048

Research Article

Complex Cepstrum Based Voice Conversion Using Radial Basis Function

Jagannath Nirmal,[1] Suprava Patnaik,[1] Mukesh Zaveri,[2] and Pramod Kachare[1]

[1] *Department of Electronics Engineering, SVNIT, Surat, India*
[2] *Department of Computer Engineering, SVNIT, Surat, India*

Correspondence should be addressed to Jagannath Nirmal; jhnirmal1975@gmail.com

Received 30 November 2013; Accepted 26 December 2013; Published 6 February 2014

Academic Editors: K. S. Chuang and A. Maier

The complex cepstrum vocoder is used to modify the speaker specific characteristics of the source speaker speech to that of the target speaker speech. The low time and high time liftering are used to split the calculated cepstrum into the vocal tract and the source excitation parameters. The obtained mixed phase vocal tract and source excitation parameters with finite impulse response preserve the phase properties of the resynthesized speech frame. The radial basis function is explored to capture the nonlinear mapping function for modifying the complex cepstrum based real and imaginary components of the vocal tract and source excitation of the speech signal. The state-of-the-art Mel cepstrum envelope and the fundamental frequency (F_0) are considered to represent the vocal tract and the source excitation of the speech frame, respectively. Radial basis function is used to capture and formulate the nonlinear relations between the Mel cepstrum envelope of the source and target speakers. Mean and standard deviation approach is employed to modify the fundamental frequency (F_0). The Mel log spectral approximation filter is used to reconstruct the speech signal from the modified Mel cepstrum envelope and fundamental frequency. A comparison of the proposed complex cepstrum based model has been made with the state-of-the-art Mel Cepstrum Envelope based voice conversion model with objective and subjective evaluations. The evaluation measures reveal that the proposed complex cepstrum based voice conversion system approximate the converted speech signal with better accuracy than the model based on the Mel cepstrum envelope based voice conversion.

1. Introduction

The voice conversion (VC) system extracts the features of the source and the target speaker sound's and formulates the mapping function to modify the features of the source speaker sound's such that the resynthesized speech sound's as if spoken by a target speaker [1]. Application of VC includes the personification of text to speech, design of multispeaker based speech synthesis system, audio dubbing, karaoke applications, security related system, the design of speaking aids for the speech impaired patient, broadcasting, and multimedia applications [2–4]. The VC involves the transformation of speaker specific characteristics such as vocal tract parameters, source excitation, and long term prosodic parameters with that of desired speaker parameters [5]. The vocal tract parameters are relatively more prominent for identifying the speaker uniqueness than the source excitation [5].

Several methods have been reported in the literature to characterize the spectrum of the speech frame, namely, Formant Frequency (FF), Formant Bandwidth (FBW) [1], Linear Predictive Coefficients (LPC) [6], Reflection Coefficients (RC) [7], Log Area Ratio (LAR) [8], Cepstrum Coefficients [9], Mel cepstrum envelope (MCEP) [10], Wavelet Transform (WT) [11], and Mel generated spectra [12]. Line Spectral Frequency (LSF) [13, 14] is a direct mathematical transformation of LPC, which has a special attraction in representing the vocal tract as it smoothly traces the shape of formants and antiformants and overcomes the interpolation, quantization, and stability issues of the LPC. However, LP related features does not assume nonstationary characteristics of the speech signal within a frame and therefore fail to analyze the local speech events accurately [15]. Further, a very accurate approach STRAIGHT [16] has also been proposed. It needs enormous computations and therefore, it is inappropriate

for real time applications. Another approach using Mel Frequency Cepstrum Coefficients (MFCC) have been proposed [17], which properly model both spectral peaks and valleys. However, the main toil of MFCC synthesis is to loose pitch and phase related information [17].

The conventional parametric speech production model like LPC, real cepstrum [18–20], and Liljencrants-Fant (LF) [21] models is based on minimum phase model with infinite impulse response [22]. In fact, a completely different category of glottal flow estimation relies on the mixed-phase model of speech [22, 23]. According to this estimation, the speech signal is composed of both maximum (i.e., anticausal) and minimum phase (i.e., causal) components. The return phase of the glottal pulse components and vocal tract impulse response is part of minimum phase signals, whereas the open phase of the glottal flow is considered as maximum phase of the signal [24]. It has been shown in the literature that the mixed phase models are appropriate for representing the voiced speech [25]. The real cepstrum with minimum phase discards the glottal flow information of speech. However, the complex cepstrum incorporates phase as glottal pulse information during speech synthesis [25]. The complex cepstrum representation of the speech signal allows noncasual modeling of short time speech frame, which is actually observed in natural speech [22–24]. Complex cepstrum perform well in speech synthesis and speech modeling [25, 26].

For the development of appropriate transformation model, various mapping functions have been proposed in the literature such as Vector Quantization (VQ) based codebook mapping [6] and Gaussian Mixture Model (GMM) based transformation models [3, 9, 10]. Fuzzy vector quantization [27] and a Speaker Transformation Algorithm using Segmental Code-book (STASC) have been proposed to overcome limitations of VQ based model [14]. In addition Dynamic Frequency Warping (DFW) [28] have also been used for transformation of the spectral envelope. The GMM oversmoothing issue is resolved via maximum likelihood estimators and hybrid methods [29]. The dynamic kernel partial least square regression technique has also implied [12] for spectral transformation. In fact, the relation between the shapes of the vocal tracts of the different speakers are highly nonlinear, to capture this nonlinearity between the vocal tracts artificial neural network has been explored in the literature [10, 11, 14, 18, 30].

In addition to vocal tract, the source excitation contains vital speaker-specific characteristics [1, 3], so it is necessary to properly modify the excitation signal to accurately synthesize the target speaker's voice [4]. Very few methods have been discussed in the literature for excitation signal transformation such as residual copying, but the converted sound seems to be a third speaker's voice [31], another method is residual prediction [3]. However, it has the problem of over smoothening. In order to alleviate the over smoothening problem of residual prediction, residual selection method, unit selection method [31], and combination of residual selection and unit selection have been also explored in the literature [32]. The Artificial Neural Network model has also applied to modify the residual signal but time domain residual transformation loses the correlation in the speech production model which leads to distortion in speech signal [12].

In this paper, the prominent complex cepstrum vocoder is employed to model the vocal tract and source excitation of the speech. The low time and high time lifters are designed to separate the complex cepstrum into vocal tract and source excitation parameters with real and imaginary components. The reasons behind the use of radial basis function (RBF) based the transformation model are its fast training ability, desirable computational efficiency, and interpolation property. The RBF based mapping function are trained separately to capture the nonlinear relations for modifying the real and the imaginary components of cepstrum based vocal tract and source excitation of the source speaker to that of the target speaker utterance's. Similarly, the MCEP parameters of source speaker's utterances are also modified according to the target speaker's utterances using RBF. The fundamental frequency between source and target speaker's utterances is modified using mean and standard deviation approach [10]. Mel log spectral approximation (MLSA) filter [33] is used to reconstruct the speech signal from modified MCEP and fundamental (F_0).

Finally, the performance of the proposed complex cepstrum based VC approach is compared with MCEP [34] based VC approach. This is done using various objective measures such as a performance index (P_{LSF}) [3], formant deviation [14, 30], and spectral distortion [14]. The commonly used subjective measures such as Mean Opinion Score (MOS) and ABX verify the quality and speaker identity of the converted speech signal.

This paper is organized as follows. Section 2 describes the complex cepstrum analysis with low time and high time lifters which are used to extract the cepstrum based features of the vocal tract and excitation based signals. Section 3 explains the proposed VC system based on complex cepstrum and the state-of-the-art MCEP based VC system. Radial basis based spectral mapping is described in Section 4. The experimental environment, database, and objective measures, such as performance index, formant deviation, spectrograph, and the perceptual tests, namely, Mean Opinion Score (MOS) and ABX, conducted with different human listeners are presented in Section 5. The last Section gives the overall conclusions of the paper.

2. Complex Cepstrum Analysis

According to the source-filter model of the human speech production system, the source signal excites the vocal tract and it generates the speech signal. The human speech is two-sided real and asymmetrical in nature. Hence, a mixed phase Finite Impulse Response (FIR) system may be realized which preserves the phase related information to give more accurate synthesized speech. From the signal processing point of view, the short time speech signal $s(n)$ can be considered as linear convolution of the source excitation $g(n)$ with the impulse function of the vocal tract $v(n)$. It can be defined as follows:

$$s(n) = v(n) * g(n). \tag{1}$$

By applying DTFT to the speech signal we obtain

$$S(\omega) = \sum_{n=-M}^{M} s(n)\, e^{-j\omega n}, \tag{2}$$

where M is the order of cepstrum, that is, number of one sided frequencies. The time domain convolution can be modeled as spectral multiplication of the vocal tract filter response $V(\omega)$ and source excitation response $G(\omega)$ giving the short time speech spectrum $S(\omega)$ as shown,

$$S(\omega) = V(\omega)\, G(\omega). \tag{3}$$

Cepstral analysis includes transforming the multiplied source excitation and vocal tract responses in the frequency domain into linear combination of the two components in the cepstral domain. The analysis of the speech signal needs to separate two components $V(\omega)$ and $G(\omega)$. In frequency domain logarithmic representation is used to linearly combine the components $V(\omega)$ and $G(\omega)$. The complex spectrum $S(\omega)$ can be rewritten by performing logarithmic compression

$$\widehat{S}(\omega) = \log S(\omega). \tag{4}$$

Therefore the log spectrum is further separated into two parts

$$\log S(\omega) = \log V(\omega) + \log G(\omega). \tag{5}$$

Thus, the log spectrum can be decomposed as addition of magnitude and phase components

$$\widehat{S}(\omega) = \log |S(\omega)| + j \arg S(\omega). \tag{6}$$

The imaginary part of the logarithmic spectrum is the unwrapped phase sequence [23]. Thus, phase information is no more ignored giving rise to a complex cepstrum. Hence comprising of a mixed phase system, with a finite impulse response (FIR) type, which is stable. The cepstrum is defined as

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{S}(\omega)\, e^{j\omega n} d\omega, \tag{7}$$

where $c(n)$ can be given as

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(\omega)|\, e^{j\omega n} d\omega + \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j(\omega + \arg(S(\omega)))n} d\omega. \tag{8}$$

The log spectral components that vary rapidly with frequency ω are denoted as a high time component $\log G(\omega)$ and the log spectral components that slowly with frequency ω are designated as a low time component $\log V(\omega)$ [20]. Here, $c(n)$ is time aliased version, therefore, $M > N$ condition avoids aliasing effect; N is total number of cepstrum samples.

Consider

$$\begin{aligned} l_l(n) &= \begin{cases} 1, & 0 \le n < L_c, \\ 0, & L_c \ge n \le N, \end{cases} \\ c_v(n) &= l_l(n)\, c(n), \\ l_h(n) &= \begin{cases} 1, & L_c \le n \le N, \\ 0, & \text{elsewhere}, \end{cases} \\ c_e(n) &= l_h(n)\, c(n), \end{aligned} \tag{9}$$

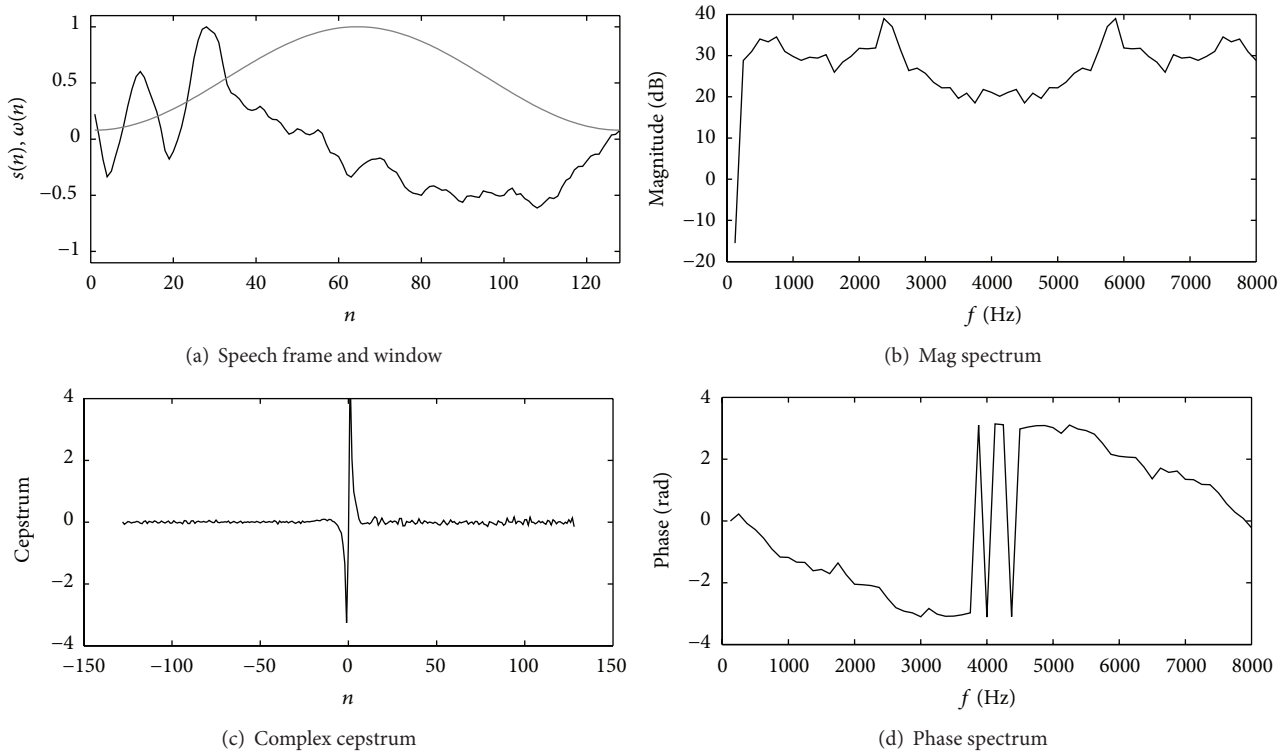
where the $c(n)$ represents complex cepstrum of speech frame, $l_l(n)$ is low time lifter, $l_h(n)$ is high time lifter. In the deconvolution stage an appropriate value of lifter index L_c is chosen to separate the two components, namely, the fast changing excitation parameter $c_e(n)$ and the slowly changing parameters, that is, vocal tract parameter $c_v(n)$. The windowed signal, the complex cepstrum with magnitude, and phase spectra are shown in Figure 1. The coefficient, $c(0)$ is the speech signal energy and the coefficients $c(n)$ for $n \ge 1$ signifies the magnitude and phase at the quefrency n in the spectrum. The vocal tract cepstrum $c_v(n)$ has coefficients with significant magnitudes at lower values of n and source excitation cepstrum; $c_e(n)$ has relatively lower magnitude coefficients for higher values of n. Theoretically, the complex cepstrum being a mixed phase results in a more accurate model of the speech signal, when compared to the minimum phase synthesis filter approach which discard the glottal flow information content in the cepstrum [18]. The cepstrum values lower than zero represents the maximum phase (i.e., anticausal) response, whereas the values above zero can be considered as the minimum phase (i.e., causal) response are shown in Figure 2. Mathematically, it can be modeled as

$$\begin{aligned} \text{Minimum Phase} &= c_{\min}(n) \\ &= \begin{cases} 0, & n = -M, \ldots, -2, -1, \\ c(n), & n = 0, \\ c(n) + c(-n) & n = 1, 2, \ldots, M, \end{cases} \\ \text{Maximum Phase} &= c_{\max}(n) = c(n) - c_{\min}(n), \\ c_{\max}(n) &= \begin{cases} c(n), & n = -M, \ldots, -2, -1, \\ 0, & n = 0, \\ -c(-n) & n = 1, 2, \ldots, M. \end{cases} \end{aligned} \tag{10}$$

The anticausal and casual cepstrum parts with the corresponding magnitude and phase spectrum are shown in Figure 3. It has been observed that the logarithmic compression involved in the cepstrum analysis helps in obtaining the mixed phase response for both voiced as well as unvoiced signals.

3. Voice Conversion Framework

In this section, the complex cepstrum based VC algorithm is proposed. The MCEP-MLSA based VC algorithm is also



(a) Speech frame and window

(b) Mag spectrum

(c) Complex cepstrum

(d) Phase spectrum

Figure 1: (a) Speech frame and window, (b) magnitude spectra, (c) complex cepstrum, and (d) phase spectra.

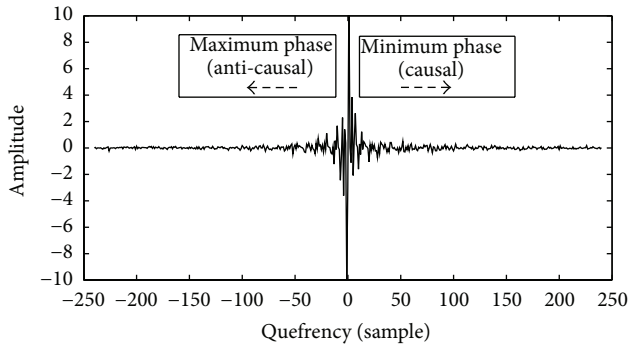


Figure 2: Complex cepstrum decomposition into maximum and minimum phase speech components.

developed for comparing the performance with the proposed algorithm.

3.1. Proposed Complex Cepstrum Vocoder Based VC. The proposed algorithm is implemented in two distinct phases: (i) training and (ii) transformation phase, as depicted in Figure 4. In the training phase, the input speech signal of the source and target speakers are normalized and silence frames are removed. The normalized speech frame is represented using homomorphic decomposition. It takes the advantages of the logarithmic scaling and the theory of convolution. The low time portion of the complex cepstrum can be approximated as a vocal tract impulse response (VT), where as high time portion of the complex cepstrum is considered as source excitation (GE) of the speech frame. The length of the rectangular lifter is chosen with regard to the accuracy of the vocal tract model and sampling frequency. Thus, the cepstrum frame is split into vocal tract impulse response and source excitation of the speech using low time and high time liftering, respectively. Even if the source and the target speaker utter the same sentence, the length of their feature vectors may be different so dynamic time warping is used to align these feature vectors. The separate RBF based mapping functions are developed for modifying the cepstrum based real and imaginary components of the vocal tract and source excitation of the source speaker according to the target speaker.

In the transformation phase followed by training phase, the parallel utterance of the test speaker speech is preprocessed to derive vocal tract and source excitation feature set based on cepstral analysis. The test feature vectors are projected to the trained RBF, in order to obtain the transformed feature vectors. The time domain features are computed by inverse transforming complex cepstrum based parameters. The modified speech frame is reconstructed by convolving the transformed vocal tract and source excitation. The similar process is adapted for all remaining frames. The overlap and add method is used to resynthesize speech from modified speech frames. Finally, the speech quality is enhanced through the postfiltering, applied to the modified speech. Figure 4 depicts the training and testing phase details of the proposed approach. The resynthesized speech from the complex cepstrum has higher perceptual quality than the speech signal constructed from the real cepstrum.

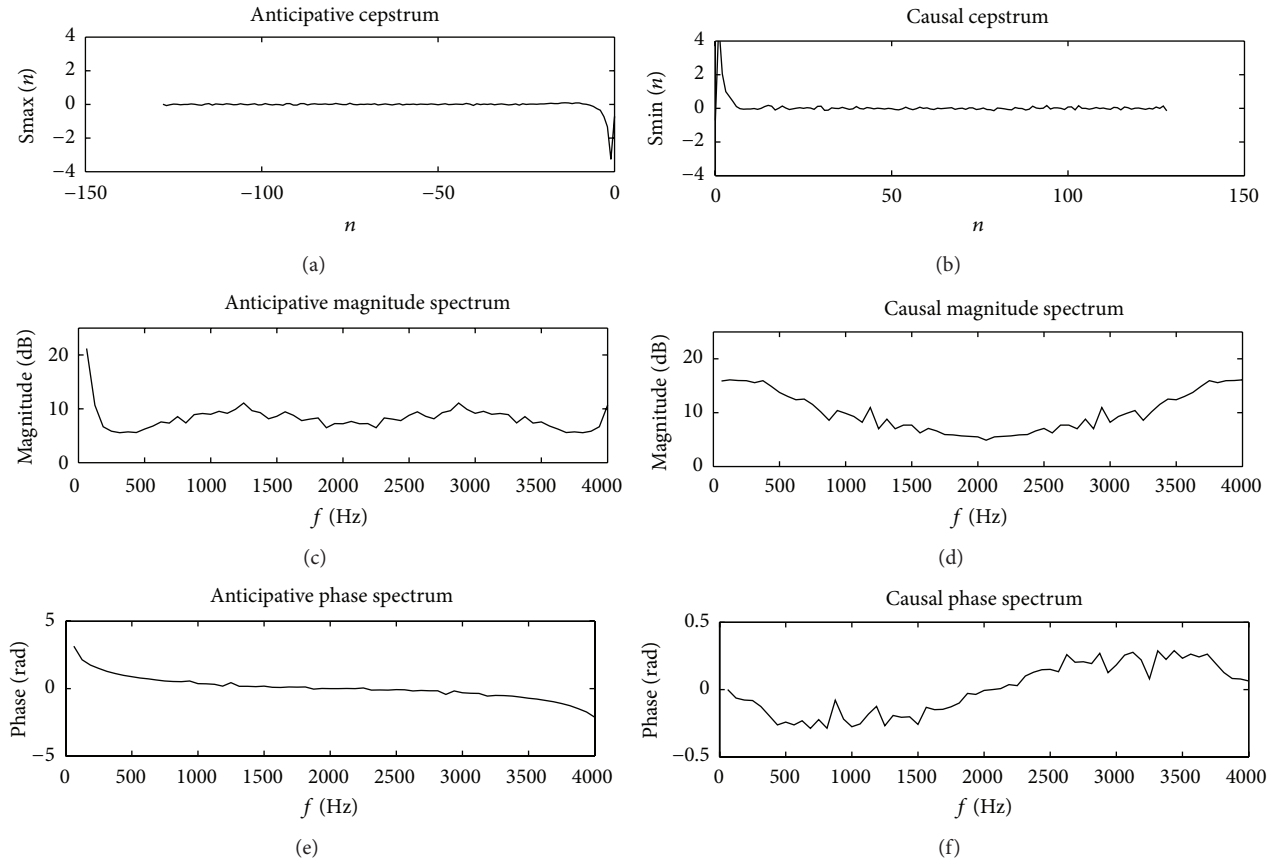


FIGURE 3: Anticausal and causal cepstrum with corresponding magnitude and phase spectrum.

3.2. Baseline Mel Cepstral Envelope Vocoder Based VC. Figure 5 depicts a block diagram of a VC system using baseline features. During the analysis step, the MCEPs are derived as spectral parameters and the fundamental frequency (F_0) is derived as excitation parameter for every 5 msec [10]. As discussed in the earlier section the feature sets obtained from the source and target speakers usually differ in time duration. Therefore, the source and target speaker's utterances are aligned using DTW. The feature set captures the joint distribution of source and target speaker using RBF to carry out VC. The excitation features (F_0) use the cepstrum method to calculate the pitch period for the frame size of 25 msec resulting into 25 MCEP features. Mean and standard deviation statistics are obtained from $\log(F_0)$ and used as feature set. In the testing phase, the parallel utterances of test speaker are used to obtain the feature vector with the procedure similar to that of the training set feature vector. In order to produce transformed feature vector, the test speaker feature vector is projected through the trained RBF model. In the synthesis stage, the transformed MCEP and F_0 are passed through the MLSA [10, 33, 35] filter. The postfiltering applied to the transformed speech signal ensures its high quality.

4. Radial Basis Function Based VC

The RBF is used to model the nonlinearity between the source and the target speaker feature vectors [11]. It is a special case of feed forward network which nonlinearly maps input space to hidden space followed by a linear mapping from a hidden space to the output space. The network represents a map from M_0 dimensional input space to N_0 dimensional output space written as $S : R_0^M \rightarrow R_0^N$. When a training dataset of input output pairs $[x_k, d_k]$; $k = 1, 2, \ldots, M_0$ is applied to the RBF model; the mapping function F is computed as

$$F_k(x) = w_{j0} + \sum_{j=1}^{m} w_{jk} \Phi\left(\left\|x - d_j\right\|\right), \tag{11}$$

where $\|\cdot\|$ is a norm usually Euclidian and computes the distance between applied input x and training data point d_j and $\Phi(\|x - d_j\|) \mid j = 1, 2 \ldots, m$ is the set of m arbitrary functions known as radial basis functions. The commonly considered form of Φ is Gaussian function defined as

$$\Phi(x) = \exp\left(\frac{\|x - \mu\|^2}{2\sigma^2}\right). \tag{12}$$

RBF neural network learning process includes training and generalized phase. The training phase constitutes the optimization of basis function parameters using input dataset to evaluate k-means algorithm in an unsupervised manner [11]. In the second phase, hidden-output neurons weight matrix is

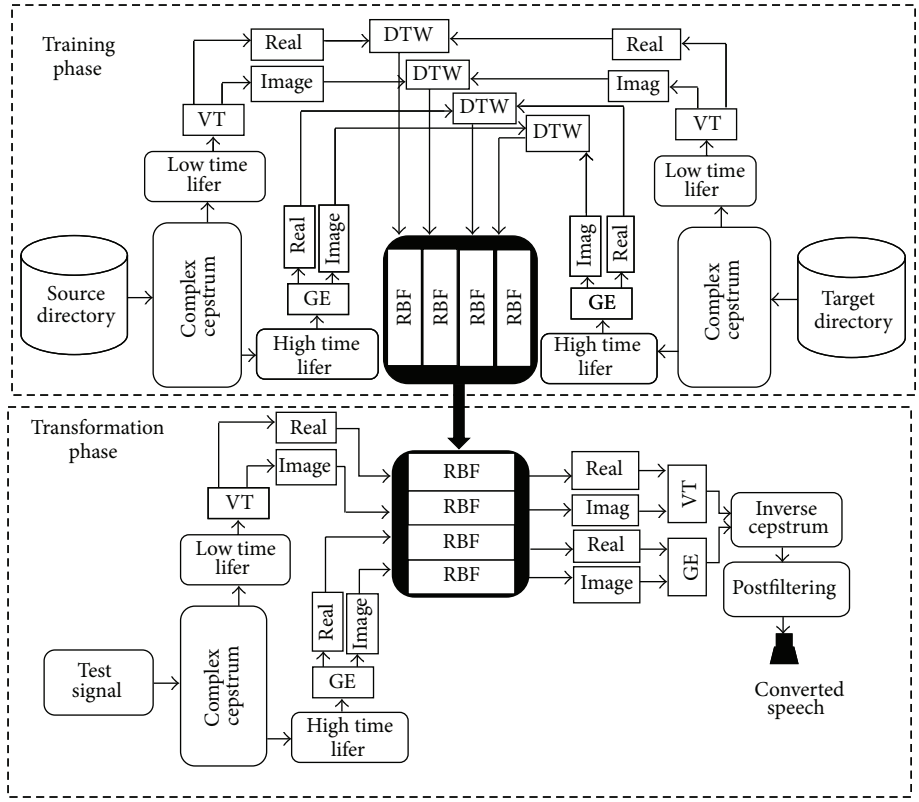


Figure 4: Functional block diagram of the complex cepstrum based VC.

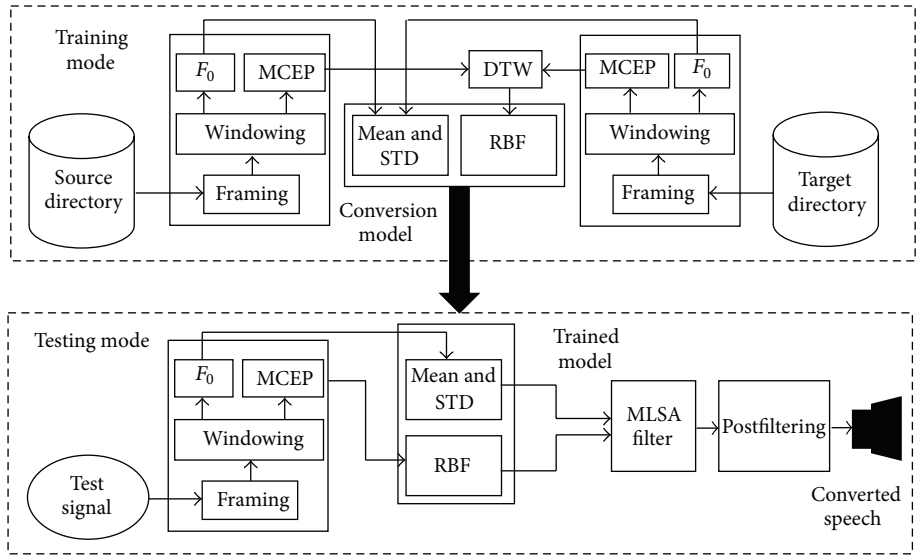


Figure 5: Block diagram of analysis and Synthesis of Mel cepstrum based VC system.

optimized by the least square sense to minimize the squared error function using the equation

$$E = \frac{1}{2}\sum_{n}\sum_{k}\left[f_k\left((x^n) - (d_k)^n\right)\right]^2, \tag{13}$$

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

$$W = \Phi^T D, \tag{14}$$

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

$$\begin{aligned} \left(\Phi^T \Phi\right) W &= \Phi^T D, \\ W &= \left(\Phi^T \Phi\right)^{-1} \Phi^T D, \end{aligned} \tag{15}$$

where $(\Phi^T\Phi)^{-1}\Phi^T$ represents the pseudoinverse 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 acoustic feature vector. The exact interpolation of RBF is acquainted with two serious problems, namely, (i) poor performance for noisy data and (ii) increased computational complexity. These problems can be addressed by modifying two RBF parameters. The first one is the spread factor which is calculated as

$$\sigma_j = 2 \times \text{avg}\left\{\left\|x - \mu_j\right\|\right\}. \tag{16}$$

The selected spread factor confirms that the individual RBFs are neither wide nor narrow. The second one is an extra bias unit which is introduced into the linear sum of activations at the desired output layer to compensate for the difference between the mean over the data set of the basis function activations and the corresponding mean of the targets. Hence, we achieve the RBF network for mapping as

$$F_k(x) = \sum_{j=0}^{m} w_{jk} \Phi\left(\left\|x - d_j\right\|\right). \tag{17}$$

In this work RBF neural networks are initialized and best networks are developed to obtain the mapping between the cepstral based acoustic parameters of the source and the target speakers. The trained networks are used to predict real and imaginary components of the vocal tract and source excitation of the target speaker's speech signal. In the baseline approach, the MCEP based feature matrices of the source and target utterances with the order of 25 are formed. Radial basis function is trained to obtain best mapping function. The best mapping function is obtained using RBF network and used to predict the MCEP parameters of the target speaker's speech signal.

5. Experimental Results

In this paper, the RBF based mapping functions are developed using CMU-ARCTIC corpus. The corpus consists of different sets of 1132 phonetically balanced parallel utterances of each speaker, sampled at 16 kHz. The corpus includes two female, that is, CLB (US Female) and SLT (US Female), and five different male such as AWB (Scottish Male), BDL (US Male), JMK (Canadian Male), RMS (US Male), and KSP (Indian Male) [36]. In this work, we have made use of the parallel utterances of the AWB (M1), CLB (F1), BDL (M2), and SLT (F2) with different speaker combinations like M1-F1, F2-M2, M1-M2 and F1-F2. For each of the speaker pairs 50 parallel sentences of source and target speakers are used for VC system training and system evaluations are made using a separate set of 25 source speaker sentences. The performance of homomorphic vocoder based VC system is compared with the state-of-the-art MCEP based VC system using different objective and subjective measures.

5.1. Objective Evaluation. The objective measures provide the mathematical analysis for determining the similarity index and quality inspection score between desired (target) and transformed speech signal. In this work, performance index, spectral distortion and formant deviation are considered as objective measures.

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, $D_{\text{LSF}}(d(n), \hat{d}(n))$ and the interspeaker spectral distortion, $D_{\text{LSF}}(d(n), s(n))$ are used for computing the P_{LSF} measure. In general, the speaker spectral distortion between signals u and v, $D_{\text{LSF}}(u, v)$ is defined as

$$D_{\text{LSF}}(u, v) = \left[\frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{P}\sum_{j=1}^{P}\left(\text{LSF}_u^{i,j} - \text{LSF}_v^{i,j}\right)^2}\right], \tag{18}$$

where N represents the number of frames, P refers to a LSF order, and $\text{LSF}_u^{i,j}$ is the jth LSF component in the frame i. The P_{LSF} measure is given as

$$P_{\text{LSF}} = \left[1 - \frac{D_{\text{LSF}}\left(d(n), \hat{d}(n)\right)}{D_{\text{LSF}}(d(n), s(n))}\right]. \tag{19}$$

The performance index $P_{\text{LSF}} = 1$ indicates that the converted signal is identical to the desired one, whereas $P_{\text{LSF}} = 1$ specifies that the converted signal is not at all similar to the desired one.

In the computation of the performance index, four different converted samples of M1 to F1, F2 to M2, F1 to F2, and M1 to M2 combinations are considered. Comparative performance between cepstrum based VC algorithm and MCEP based VC is shown in Table 1. The results specified that the performance of the complex cepstrum based VC performed better than MCEP based VC algorithm.

Along with performance index, the different objective measures, namely, deviation (D_i), root mean square error (RMSE), and correlation coefficients ($\sigma_{x,y}$), are also calculated for different speaker pairs. Deviation parameter is defined as the percentage variation in the actual (x_k) and predicted (y_k) formant frequencies derived from the speech frames. It corresponds to the percentage of test frames within a specified deviation. Deviation (D_k) is calculated as

$$D_k = \frac{|x_k - y_k|}{x_k} \times 100. \tag{20}$$

Table 1: The performance index of complex cepstrum based VC and MCEP based VC.

Type of conversion	Performance index							
	Sample 1		Sample 2		Sample 3		Sample 4	
	Cep. based VC	MCEP based VC	Cep. based VC	MCEP based VC	Cep. based VC	MCEP based VC	Cep. based VC	MCEP based VC
M1-F1	0.7679	0.6230	0.7483	0.6356	0.7127	0.6080	0.8350	0.6768
F2-M2	0.7389	0.5781	0.7150	0.6988	0.7780	0.6908	0.7848	0.6845
F1-F2	0.7921	0.6576	0.6908	0.5954	0.6740	0.6209	0.7946	0.6925
M1-M2	0.7023	0.6490	0.6821	0.6012	0.6432	0.5801	0.7852	0.7012

Table 2: Prediction performance of MCEP based for formant frequencies.

Transformation model	Formant frequencies	% Predicted frame within deviation								
		2%	5%	10%	15%	20%	25%	50%	μ_{RMSE}	$\Upsilon_{X,y}$
M1-F1	F1	51	74	80	81	83	85	90	4.45	0.7235
	F2	45	63	68	78	82	87	89	3.73	0.8182
	F3	57	62	79	86	87	89	92	3.34	0.8703
	F4	69	79	84	89	88	90	100	2.39	0.8629
F2-M2	F1	36	58	67	74	82	86	90	4.28	0.7190
	F2	57	82	86	87	87	89	91	6.30	0.7238
	F3	72	77	89	91	92	94	95	5.23	0.7474
	F4	66	74	89	90	93	95	100	4.91	0.7957

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

$$\mu_{\text{RMSE}} = \frac{\sqrt{\sum_k |x_k - y_k|^2}}{\overline{x}} \times 100,$$
$$\sigma = \sqrt{\sum_k d_k^2}, \quad d_k = e_k - \mu, \tag{21}$$
$$e_k = x_k - y_k, \quad \mu = \frac{\sum_k |x_k - y_k|}{N}.$$

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. The parameter d_k is the error in the deviation. The correlation coefficient $\Upsilon_{X,y}$ is the parameter which is to be determined from the 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 $\Upsilon_{X,y}$ and $\text{COV}(X, Y)$ are calculated using

$$\Upsilon_{X,y} = \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y},$$
$$\text{COV}(X, Y) = \frac{\sum_k |(x_k - \overline{x})(y_k - \overline{y})|}{N}. \tag{22}$$

The objective measures, namely, deviation (D_i), root mean square error (RMSE), and correlation coefficients ($\Upsilon_{X,y}$) of M1-F1 and F2-M2 are obtained for MCEP based VC algorithm and shown in Table 2. Similarly, the Table 3 shows the measures obtained for proposed VC system. From the tables it can be observed that the μ_{RMSE} between the desired and the predicted acoustic space parameters for proposed model are 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. The scatter plots for first, second, third, and fourth formant frequencies for MCEP based VC and complex cepstrum based VC models are shown in Figures 6 and 7, respectively. Figures show that complex cepstrum VC based vocal tract envelope in term of predicted formants closely orient towards the desired speech frames formants as compared to MCEP based predicted formants. The clusters obtained using complex cepstrum based VC are more compact and diagonally oriented than that using MCEP based VC. As perfect prediction means all the data points in scatter plot are diagonally oriented in right side. The compact clusters obtained for proposed method implies its ability to capture the formant structure of desired speaker

The transformed formant patterns for a specific frame of source and target speech signal are obtained using both complex cepstrum and MCEP based VC models and shown in Figures 8(a) and 8(b), respectively. Figure 8(a) depicts that the patterns of particular target signal closely follows the corresponding transformed signal, whereas Figure 8(b) shows that the predicted formant pattern closely follows the target pattern only for lower formants.

Figure 9(a) shows the normalized frequency spectrogram of desired and transformed speech signals obtained from M1 to F1 and F2 to M2 of complex cepstrum based VC model. Similarly, Figure 9(b) shows the spectrogram for M1 to F1 and

Table 3: Prediction performance of complex cepstrum based for formant frequencies.

Transformation model	Formant frequencies	% Predicted frame within deviation								
		2%	5%	10%	15%	20%	25%	50%	μ_{RMSE}	$\Upsilon_{X,y}$
M1-F1	F1	59	80	89	91	91	93	95	4.45	0.7197
	F2	52	72	85	88	91	92	95	3.55	0.8149
	F3	63	83	90	92	93	95	99	2.65	0.8837
	F4	72	86	91	93	95	97	100	2.049	0.8909
F2-M2	F1	38	60	71	78	80	86	90	8.56	0.757
	F2	60	82	89	92	92	93	98	3.62	0.790
	F3	72	87	92	95	95	95	100	2.93	0.778
	F4	70	86	91	96	97	99	100	2.41	0.756

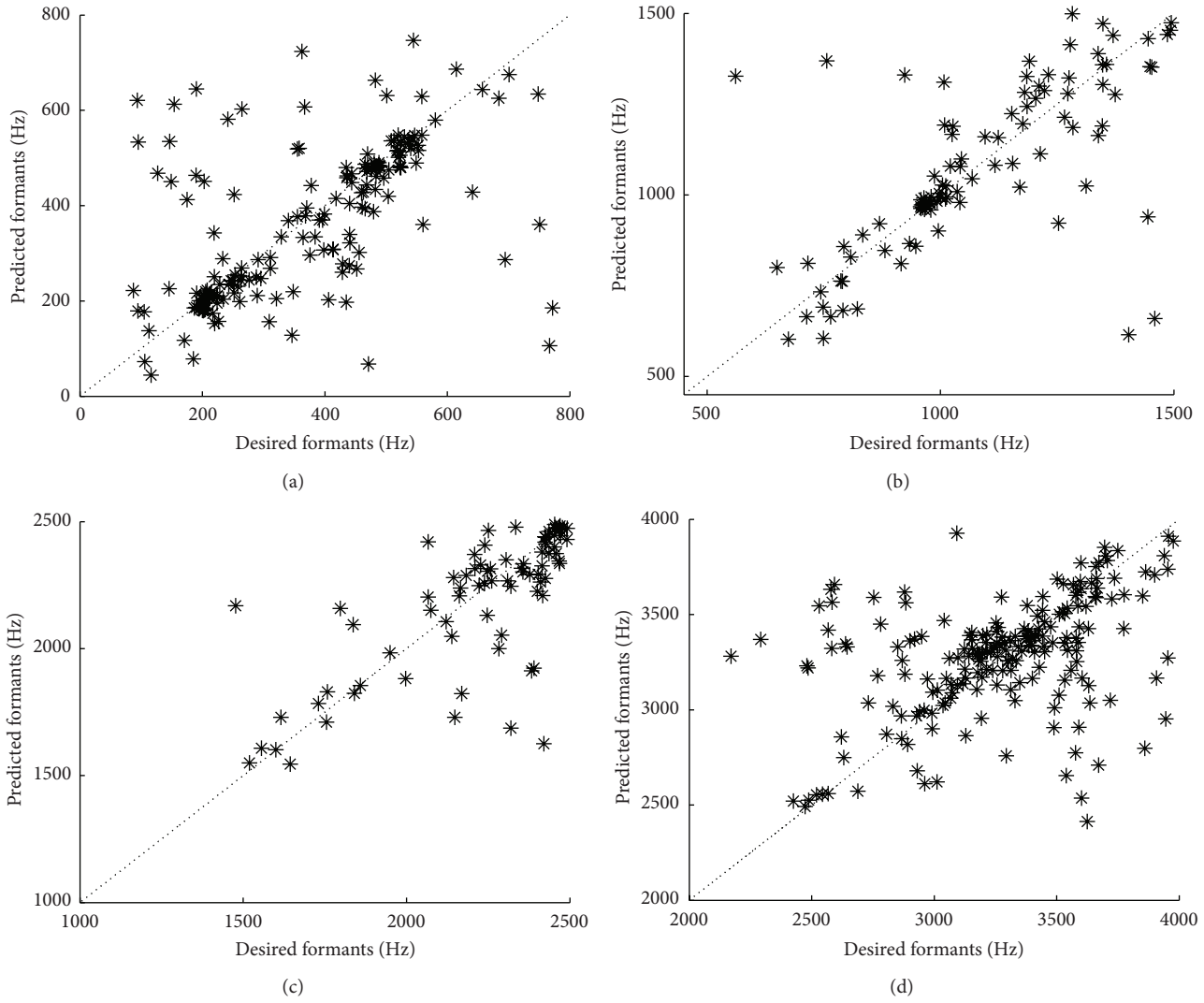


Figure 6: Desired and predicted formant frequencies for F2 to M2 VC using MCEP based approach (a) first formant, (b) second formant, (c) third formant, and (d) fourth formant.

F2 to M2 for the MCEP based VC model. It has been observed that the dynamics of the first three formant frequencies in both the algorithms are closely followed in the target and the transformed speech samples.

5.2. Subjective Evaluation. The effectiveness of the algorithm is also evaluated using listening tests. These subjective tests are used to determine the closeness between the transformed and target speech sample. The mapping functions are developed using 50 parallel utterances of the source and target speakers. Twenty-five different synthesized speech utterances are obtained from the mapping function for inter- and intragender speech conversion and corresponding target utterances are presented to twelve listeners. They are asked to

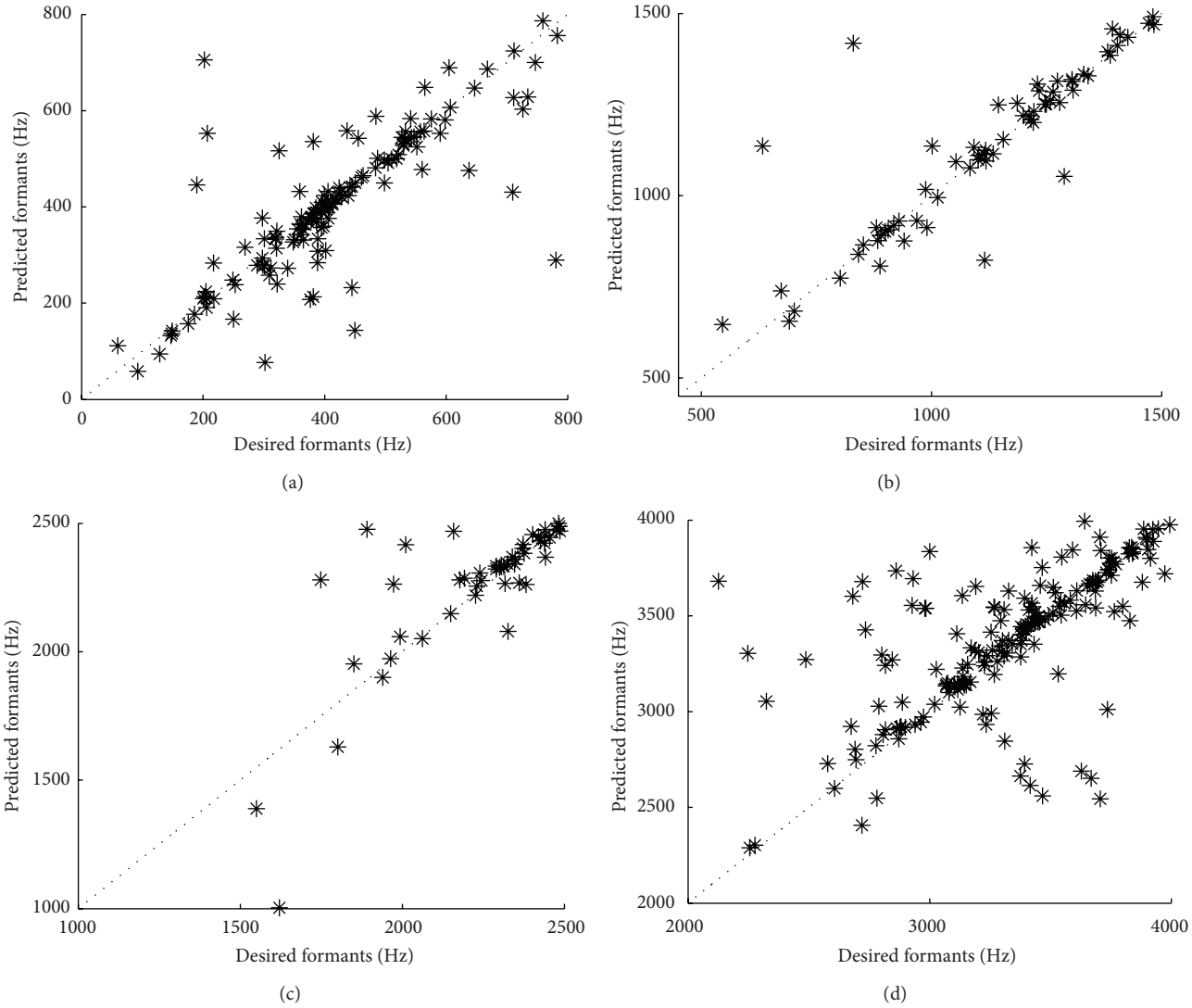


Figure 7: Desired and predicted formant frequencies for F2 to M2 VC using the complex cepstrum based (a) first formant, (b) second formant, (c) third formant, and (d) fourth formant.

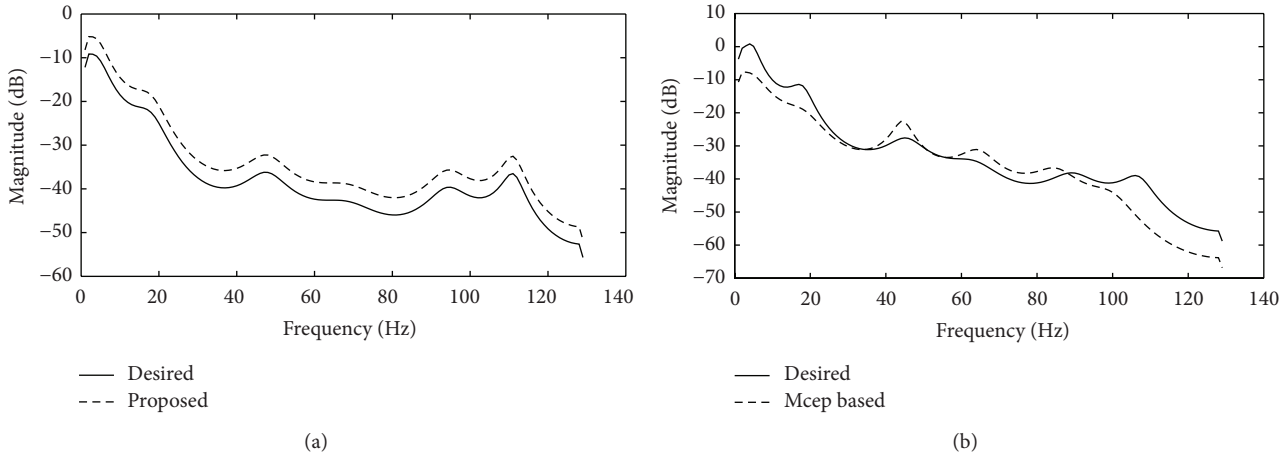


Figure 8: Target and transformed spectral envelopes of the desired speaker using (a) complex cepstrum based VC and (b) MCEP based VC.

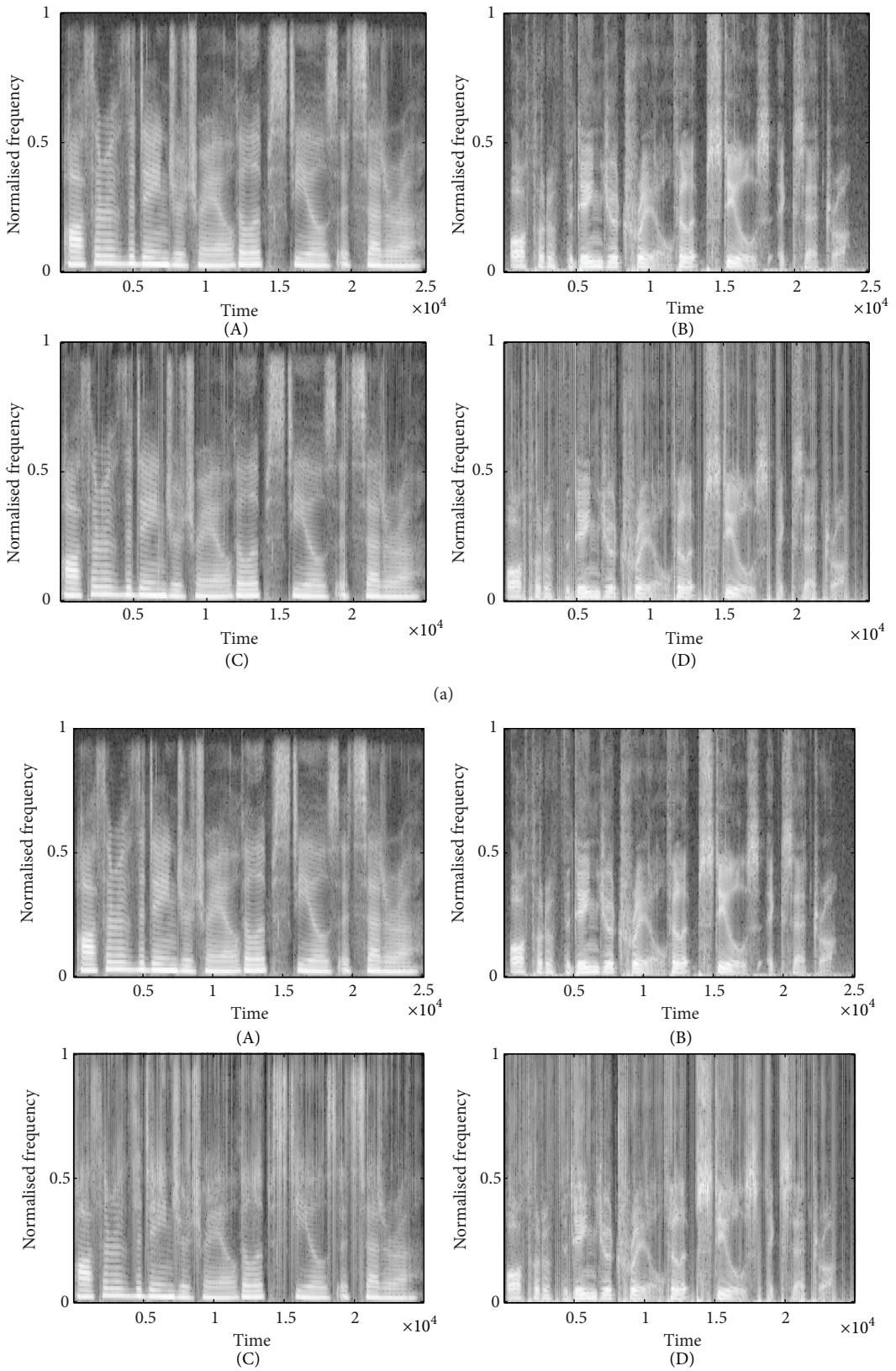


Figure 9: Spectrogram of the desired and the transformed signal for M1 to F1 ((A) to (C)) and F2 to M2 ((B) to (D)) using (a) complex cepstrum based VC and (b) MCEP based VC.

Table 4: MOS and ABX evaluations of complex cepstrum and MCEP based VC models.

Conversion data	MOS		ABX	
	Cepstrum based	MCEP based	Cepstrum based	MCEP based
M1-F1	4.64	4.31	4.55	4.25
F2-M1	4.18	3.92	4.34	4.23
M1-M2	4.07	3.88	4.19	4.06
F1-F2	4.24	3.76	4.36	4.13

evaluate their relative performance in term of voice quality (MOS) and speaker identity (ABX) with corresponding source and target speaker speech samples on a scale of 1 to 5, where rating 5 specifies an excellent match between the transformed and target utterances, rating 1 indicates a poor match, and the other ratings indicate different levels of variation between 1 and 5. The ratings given to each set of utterances are used to calculate the MOS for different speaker combinations like M1 to F1, M1 to M2, F1 to F2, and F2 to M2; the results are presented in Table 4. The dissimilarity in the length of the vocal tract and the intonation patterns of different genders is the major reason for variation in the MOS results for source and target utterances of different genders. The ABX (A: Source, B: Target, X: Transformed speech signal) test is also performed using the same set of utterances and speakers. In the ABX test, the listeners are asked to judge whether the unknown speech sample X sounds closer to the reference sample A or B. The ABX is a measure of identity transformation. The higher value of ABX percentage indicates that the transformed speech lies in close proximity of the target utterance. The results of the ABX test are also shown in Table 4.

6. Conclusion

The VC algorithm comprising of complex cepstrum, that preserves the phase related information content of the synthesized speech outcome, is presented. A mixed phase system is designed to yield far better transformed speech signal than the minimum phase systems. The vocal tract and excitation parameters of the speech signal are obtained with the help of low and high time liftering. Radial basis functions are explored to capture the nonlinear mapping function for modifying the real and imaginary parts of the vocal tract and source excitations of the source speaker speech to that of the target speaker speech. In baseline VC algorithm MCEP method is used to interpret the vocal tract whereas, the fundamental frequency (F_0) represent the source excitation. The RBF based mapping function is used to capture the nonlinear relationship between the MCEP of the source speaker to that of the target speaker and statistical mean and standard deviation is used for transformation of fundamental frequency. The proposed complex cepstrum based VC is compared with the MCEP based VC using various objective and subjective measures. The evaluation results reveal that the complex cepstrum based VC performs slightly better than the MCEP based VC model in term of speech quality and speaker identity. The reason may be the fluctuation of MLSA filter parameters with limited margins in Padé approximation. It may be unstable momentarily, when the parameters vary rapidly by contrast the complex cepstrum with finite impulse response is always stable.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] H. Kuwabara and Y. Sagisak, "Acoustic characteristics of speaker individuality: control and conversion," *Speech Communication*, vol. 16, no. 2, pp. 165–173, 1995.

[2] K.-S. Lee, "Statistical approach for voice personality transformation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 641–651, 2007.

[3] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 813–816, May 2001.

[4] D. Sundermann, "Voice conversion: state-of-the-art and future work," in *Proceedings of the 31st German Annual Conference on Acoustics (DAGA '01)*, Munich, Germany, 2001.

[5] D. G. Childers, B. Yegnanarayana, and W. Ke, "Voice conversion: factors responsible for quality," in *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '85)*, vol. 1, pp. 748–751, Tampa, Fla, USA, 1985.

[6] M. Abe, S. Nakanura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 655–658, 1988.

[7] W. Verhelst and J. Mertens, "Voice conversion using partitions of spectra feature space," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, pp. 365–368, May 1996.

[8] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Communication*, vol. 16, no. 2, pp. 139–151, 1995.

[9] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[10] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[11] C. Orphanidou, I. M. Moroz, and S. J. Roberts, "Wavelet-based voice morphing," *WSEAS Journal of Systems*, vol. 10, no. 3, pp. 3297–3302, 2004.

[12] E. Helander, T. Virtanen, N. Jani, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transcation on Audio, Speech, Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[13] L. M. Arslan, "Speaker transformation algorithm using segmental codebooks (STASC)," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[14] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Computer Speech and Language*, vol. 24, no. 3, pp. 474–494, 2010.

[15] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, pp. 137–140, Adelaide, Australia, 1994.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[17] R. J. McAulay and T. F. Quatieri, "Phase modelling and its application to sinusoidal transform coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, vol. 1, pp. 1713–1716, Tokyo, Japan, 1986.

[18] J. Nirmal, P. Kachare, S. Patnaik, and M. Zaveri, "Cepstrum liftering based voice conversion using RBF and GMM," in *Proceedings of the IEEE International Conference on Communications and Signal Processing (ICCSP '13)*, pp. 570–575, April 2013.

[19] A. V. Oppenheim, "Speech analysis and synthesis system based on homomorphic filtering," *The Journal of the Acoustical Society of America*, vol. 45, no. 2, pp. 458–465, 1969.

[20] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 43–51, 1986.

[21] H. Deng, R. K. Ward, M. P. Beddoes, and M. Hodgson, "A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 445–455, 2006.

[22] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 116–119, Brighton, UK, September 2009.

[23] T. F. Quatieri Jr., "Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 328–335, 1979.

[24] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[25] M. Vondra and R. Vích, "Speechmodeling using the complex cepstrum," in *Toward Autonomous, Adaptive, and Context-Aware Ultimodal Interfaces: Theoretical and Practical, Issues*, vol. 6456 of *Lecture Notes in Computer Science*, pp. 324–330, 2011.

[26] R. Maia, M. Akamine, and M. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 4581–4584, 2012.

[27] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 594–597, June 1991.

[28] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of straight spectrum," in *Proceedings of the IEEE Interntional Conference on Acoustics, Speech, and Signal Processing*, pp. 841–844, May 2001.

[29] H. Ye and S. Young, "High quality voice morphing," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I9–I12, May 2004.

[30] R. Laskar, K. Banerjee, F. Talukdar, and K. Sreenivasa Rao, "A pitch synchronous approach to design voice conversion system using source-filter correlation," *International Journal of Speech Technology*, vol. 15, pp. 419–431, 2012.

[31] D. Sündermann, A. Bonafonte, H. Ney, and H. Höge, "A study on residual prediction techniques for voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. I13–I16, March 2005.

[32] K. S. Rao, R. H. Laskar, and S. G. Koolagudi, "Voice transformation by mapping the features at syllable level," in *Pattern Recognition and Machine Intelligence*, vol. 4815 of *Lecture Notes in Computer Science*, pp. 479–486, Springer, 2007.

[33] http://sp-tk.sourceforge.net/.

[34] S. Imai, "Cepstral analysis and synthesis on the mel-frequency scale," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '83)*, pp. 93–96, Boston, Mass, USA, 1983.

[35] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis—a unified approach to speech spectral estimation," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '94)*, pp. 1043–1046, 1994.

[36] J. Kominek and A. W. Black, "CMU ARCTIC speech databases," in *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 223–224, Pittsburgh, Pa, USA, June 2004.