Cepstrum Liftering based Voice Conversion using RBF and GMM

Jagannath Nirmal, Pramod Kachare, Suprava Patnaik, Mukesh Zaveri

***Abstract*—Voice Conversion is a technique which morphs the speaker dependent acoustical cues of the source speaker to those of the target speaker. Speaker dependent acoustical cues are characterized at different levels such as shape of vocal tract, glottal excitation and long term prosodic parameters. In this paper, low time and high time liftering is applied to the cepstrum to separate the vocal tract and glottal excitation of the speech signal. The Radial Basis Function and Gaussian Mixture Model are developed to capture the mapping functions for modifying the cepstrum based vocal tract and glottal excitation of the source speaker according to a target speaker. The subjective and Objective measures are used to evaluate the comparative performance of RBF and GMM based voice conversion system. Results indicate that the RBF based transformation can be used as an alternative to GMM based model. Subjective evaluation illustrate that the proposed algorithm maintains target voice individuality, quality of the speech signal.**

***Index Terms*— Cepstrum, gaussian mixture model, liftering, radial basis function, voice conversion.**

I. Introduction

Voice conversion modifies the speaker dependent parameters of the source speaker to make it perceptually similar to that of the target speaker. Fundamentally, voice conversion involves the transformation of vocal tract parameters, glottal excitation and long term prosodic parameters onto another speaker [11].The vocal tract parameters are relatively more prominent in specifying speaker individuality of the speech signal as compared to the glottal excitation [7] .The pitch contour is an important cue of the speaker individuality[7]. Voice conversion finds many applications, including the customization of text to speech, personification of speech, design of multi-speaker based speech synthesis system, audio dubbing, the design of speaking aids for the speech impairment patient, broadcasting and multimedia applications[1][2][3]. The Vocal tract parameters act as key parameters for voice conversion.

Jagannath Nirmal and Pramod Kachare are working as a faculty in Department of Electronics, K J Somaiya College of Engineering, Mumbai, India. (Email: - jhnirmal1975@gmail.com, pramod_1991@yahoo.com).

Suparava Patnaik is working as faculty in Department of Electronics and Telecommunication, Sardar Vallabbhai National Institute of Technology, Surat, India.(email:- ssp@eced.svnit.ac.in)

Mukesh Zaveri is working as faculty in Department of Computer Engineering, Sardar Vallabbhai National Institute of Technology, Surat, India. (email:- mazaveri@coed.svnit.ac.in)

Various techniques have been proposed in the literature to characterize the vocal tract parameters of the speech frame in the form of formant frequency, formant bandwidth [7] and linear predictive related spectral features such as Linear Predictive Coefficients (LPC) [8], Reflection Coefficients [9], Log Area Ratio [10], Line Spectral Frequencies (LSF) [1][11]. Some of the cepstrum based features like Mel Frequency Cepstrum Coefficients (MFCC) [13], Mel Cepstrum Envelope (MCEP) [12] and wavelet transform [4] also describe vocal tract features to a good extent. Kawahara et. al. proposed very accurate model STRAIGHT, still it requires considerable amount of computation and hence is not appropriate for real time applications [7].

Several techniques have been proposed to develop the mapping function which approximates transformation from source feature vectors to that of the target speaker feature vectors. Such techniques are based on Vector Quantization [8], Codebook Mapping [1], Dynamic Frequency Warping and Interpolating [14], regression model [16], Gaussian Mixture Model [13][14] and Hidden Markov Model [15]. To capture the non-linearity between shape of the vocal tract of source and target speaker Artificial Neural Network (ANN) based approach is proposed [6][11][12].

The contributions of this paper are, i) the Cepstrum based approach have been proposed to separate the vocal tract and glottal excitation of the speech signal. ii) Low time liftering is applied to the cepstrum for extracting the vocal tract parameters whereas high time lifter separates the glottal excitation of the speech frame. iii) The strong generalization ability of Radial Basis Function is used to map the cepstrum based vocal tract and glottal excitation of the source speaker according to the target speaker. iv) The state of art Gaussian Mixture Model (GMM) mapping rules are developed for the proposed voice conversion system. v) Subjective and objective measures are used to evaluate the comparative performance of the RBF and GMM based voice conversion systems.

The rest of the paper is structured as follows: next section describes the acoustical feature extraction technique. Section three provides detailed sketch of the proposed algorithm. Development of spectral mapping models using RBF and GMM is explained in sections four and five respectively. The experimental results of subjective and objective measures are conveyed in section six. The last section derives the conclusion of the paper.

978-1-4673-4866-9/13/$31.00 ©2013 IEEE

II. FEATURE EXTRACTION

Voice conversion is generally carried out using speech analysis and synthesis systems [12]. Paper focuses on homomorphic analysis and synthesis technique as the splitting up the speech signal in the vocal tract and glottal excitation parameters. The input speech signal is pre-processed and segmented in 30msec frame with 50 % (i.e. 15msec) overlapping frames. Each frame is multiplied by hamming window smoothens the signal and removes artefacts that will be generated during reconstruction. The short time speech signal can be modelled as linear convolution of the source excitation and the impulse function of vocal tract. Mathematically,

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

where, $s(n)$ is speech signal, $v(n)$ is vocal tract impulse function and $g(n)$ is signal excitation. The time domain convolution can be modelled as spectral multiplication of the shape of vocal tract $v(w)$ and source glottal excitation $g(w)$ giving the short time speech signal $s(w)$ shown as,

$$s(w) = V(w)\, G(w) \tag{2}$$

The logarithm of the above equation is:

$$log|s(w)| = log|v(w)| + log|g(w)| \tag{3}$$

Finally, cepstrum is computed by taking the inverse Fourier transform of equation log magnitude spectrum. The objective of the cepstrum analysis is to separate the vocal tract and glottal excitation. This logarithm spectrum is separated as two parts namely, the log spectral components that vary rapidly with frequency w denoted as high-time components *G(w)* and the log spectral components that vary slowly with frequency w designated as low-time components *V(w).*

$$L_l(n) = \begin{cases} 1, & 0 \le n < L_c \\ 0, & L_c \ge n \le \frac{N}{2} \end{cases} \tag{4}$$

$$C_v(n) = L_l(n)c(n) \tag{5}$$

$$L_h(n) = \begin{cases} 1, & L_c \le n \le \frac{N}{2} \\ 0, & elsewhere \end{cases} \tag{6}$$

$$C_e(n) = L_h(n)c(n) \tag{7}$$

Where, $c(n)$ is represented as a cepstrum of speech frame, $L_l(n)$ is low time lifter, $L_h(n)$ is high time lifter, $C_v(n)$ is the cepstrum based vocal tract parameters, $C_e(n)$ is cepstrum based glottal excitation and N is total no of cepstrum samples . Hence using an appropriate lifter, we can separate the two components namely, the excitation parameters and the vocal tract parameters, as shown in figure 1. This process is called de-convolution.

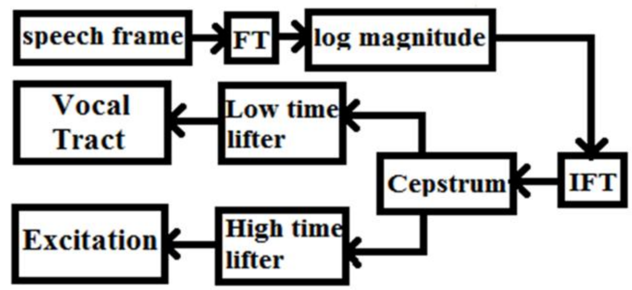
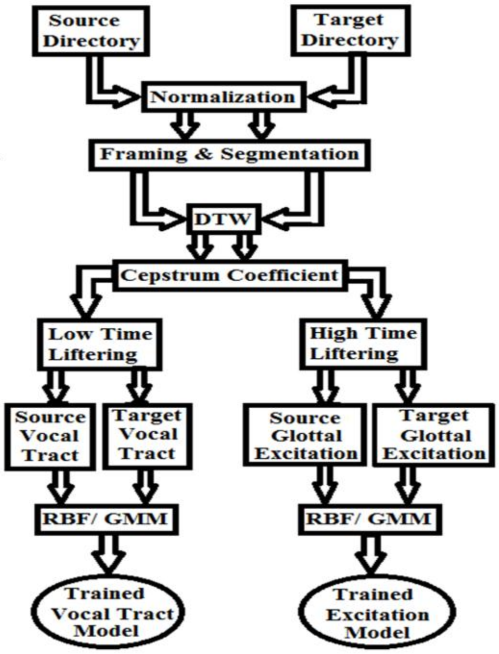


Fig. 1. Flow Diagram for Cepstrum based feature Extraction

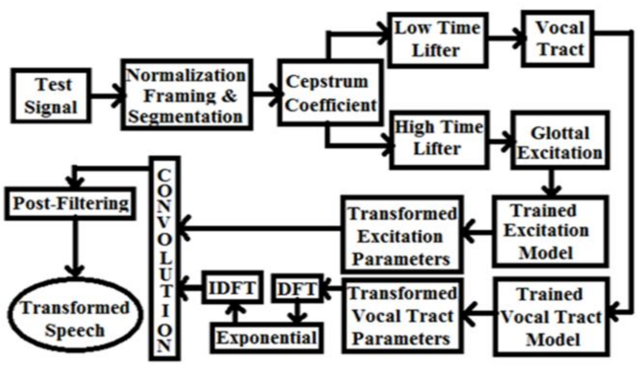
III. PROPOSED ALGORITHM

The proposed algorithm is carried out in two phases, I) training and II) transformation phase; as depicted in figure 2. In the training phase, the input speech signal is normalized and silence frames are removed from the phonetically balanced utterances of source and target speaker speech signals using Voice Activity Detection (VAD). After normalization of the voice activated signal, it is further pre-processed (as discussed in section 2) and aligned using dynamic time warping [12]. The cepstrum is computed as the inverse Fourier transform of the log magnitude Fourier transform of each speech the aligned frames [17]. The low time portion of the cepstrum can be approximated as a vocal tract impulse response whereas high time portion of the cepstrum is considered as glottal excitation of the speech frame [17]. To separate the vocal tract impulse response and glottal excitation of the speech frame, the low time and high time lifter are designed. The cepstrum frame is passed through the low time lifter to extract vocal tract impulse response, whereas the glottal excitation is obtained by passing the cepstrum through the high time lifter. The RBF and GMM based models were explored to capture the mapping function for modification of the cepstrum based vocal tract and glottal excitation of the source speaker according to the target speaker as explained in figure 2 (a). The RBF and GMM based mapping functions are explained in the following section. The training phase is followed by the transformation phase.

In transformation phase, the parallel utterance of the test speaker speech is pre-processed to obtain cepstrum based vocal tract and glottal excitation feature set. The test feature vector is projected to the trained RBF and GMM, in order to obtain the transformed coefficients. As per the transformation figure 2 (b) the time domain vocal tract impulse function is computed by taking the inverse Fourier transform of exponential of the Fourier transform of cepstrum based vocal tract parameters. To reconstruct the target speech, the time domain vocal tract parameters are convolved with glottal excitation. Overlap and add method is used to reconstruct the synthesized speech. The converted speech signal is passed through the post filtering block. The similar process is adapted for all remaining samples. Figure 2(a) depicts the training phase and Figure 2(b) represents the testing phase or transformation phase.



(a) Training Phase



(b) Transformation Phase

Fig. 2. Proposed Cepstrum based Voice Conversion Algorithm.

IV. RBF BASED VOICE CONVERSION

For developing the mapping function between source and target cepstrum based vocal tract a RBF network is used. RBF based model capture globally non-linear mapping function in the framework of voice conversion [4]. 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 [x_k, d_k]; k=1, 2……..M_0 is presented to the RBFNN model the mapping function F is computed as [5],

$$F_k(x) = \sum_{j=1}^{m} w_{jk}\ \Phi\left(\left\|x - d_j\right\|\right) \tag{8}$$

where, $\|.\|$ is a norm usually Euclidian, computes the distance between applied input x and training data point d_j .Above equation can also be written in matrix form as [4][18],

$$F(x) = W\ \Phi \tag{9}$$

where, $\Phi\left(\left\|x - d_j\right\|\right)|$j =1, 2,….., m is the set of m arbitrary functions known as Radial Basis Functions. The commonly considered form of Φ is Gaussian function defined as [4][18],

$$\Phi(x) = exp(\|x - \mu\|^2/2\sigma^2) \tag{10}$$

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 neuron's weights are optimized in a least square sense by minimizing squared error function,

$$E = \frac{1}{2}\sum_n \sum_k [f_k(x^n) - (d_k)^n)]^2 \tag{11}$$

where $(d_k)^n$ is desired value for k^{th} output unit when input to the network isx^n. The weight vector is determined as,

$$W= \Phi^{\,T} D \tag{12}$$

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

$$(\Phi^T\Phi\ )W=\Phi^{\,T} D \tag{13}$$

$$W=(\Phi^T\Phi\ )^{-1}\Phi^{\,T} D \tag{14}$$

where, $(\Phi^T\Phi\ )^{-1}\Phi^{\ T}$ 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. The performance of RBFNN can be tuned by optimizing the kernel parameters like kernel centers and spread factors which results into high quality voice conversion. In our work, we have calculated spectral distortion [14] for different kernel spread factors and hidden neurons as shown in figure 3. We have selected the spread factor of 0.01 with lowest spectral distortion.

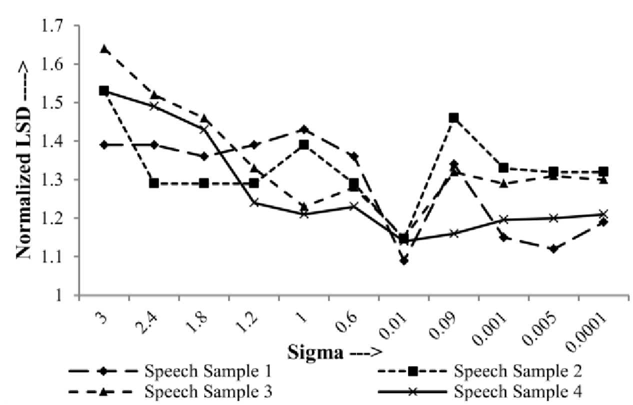


Fig. 3. Performance evaluation of RBF network for various spread factors

By selecting optimum sigma and goal the optimum network is modeled during training phase. The optimum network derived during training phase is used to predict the target speaker low time cepstrum acting as vocal tract and high time cepstrum represents the glottal excitation.

V. GMM BASED VOICE CONVERSION

The GMM is a soft decision classifier where each class has a Gaussian distribution, The GMM assumes that the probability distribution of the observed parameters takes the following parametric form [2].

$$p(x) = \sum_{i=1}^{m} \alpha_i N(x, \mu_i \ \Sigma i) \qquad (15)$$

where, m is the number of mixture models $N(x, \mu, \Sigma)$ denotes the p dimension distribution with mean (μ_i) and covariance matrix (Σ_i) defined by

$$N(x, \mu, \Sigma) = \frac{1}{\sqrt{2\pi^p \ \Sigma}} e^{\frac{1}{2}(z-\mu)^T \Sigma^{-1} (Z-\mu)} \qquad (16)$$

The weighting factor α_i is prior probability of class i with constraints $\sum_{i=1}^{m} \alpha_i = 1 \quad \alpha_i \geq 1$.

The input vector x_i is assumed to be independent. The conditional probability that a given observation vector x belongs to the components C_i of the GMM is given by Bayes rule.

$$p\left(\frac{ci}{x}\right) = \frac{\alpha_i N(x, \mu_i \ \Sigma_i)}{\sum_{j=1}^{m} \alpha_j N(x_j \mu_j \Sigma_j)} \qquad (17)$$

The parameters α, μ and Σ can be estimated with the expectation maximization (EM) algorithm. The EM algorithm is a unsupervised learning in which the component information is unavailable, The EM algorithm is find the parameters in the GMM that gives maximum likelihood of observed data X. The parameters of the conversion function is estimated by joint density of source and target features, A joint vector Z= $[x^T \ y^T]^T$ where x and y are the aligned source and target speech feature vector is used to estimate GMM parameters the following parametric form is assumed for the conversion function [2].

$$F(x) = \sum_{q=1}^{m} p\left(\frac{c_q}{x}\right) [\mu_q^y \sum_q^{yx} \Sigma_q^{xx^{-1}} (x - \mu_q)^x] \qquad (18)$$

where, p (c_q/x) is the conditional probability

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{xx} & \Sigma_q^{xy} \\ \Sigma_q^{yx} & \Sigma_q^{yy} \end{bmatrix} \qquad (19)$$

$$\mu_q = \begin{bmatrix} \mu_q^x \\ \mu_q^y \end{bmatrix} \qquad (20)$$

where, m is the number of GMM mixtures in voice conversion. The GMM joint density based model is used as a baseline model for vocal tract mapping between the source and target speakers [2]. The GMM based method effectively converts the acoustical cues of the speech signal; the quality of the converted speech is over scaled by excessive smoothing of the converted spectra. For cepstrum based GMM based mapping, the number of mixtures varies from 2 to 64 dependent on the amount of training data, however it has been observed in below figure 4 that, the 64 number of mixtures can characterized the acoustic space of the speaker. The trained transformation model developed during the training is used to predict the cepstrum based vocal tract and glottal excitation of the target.

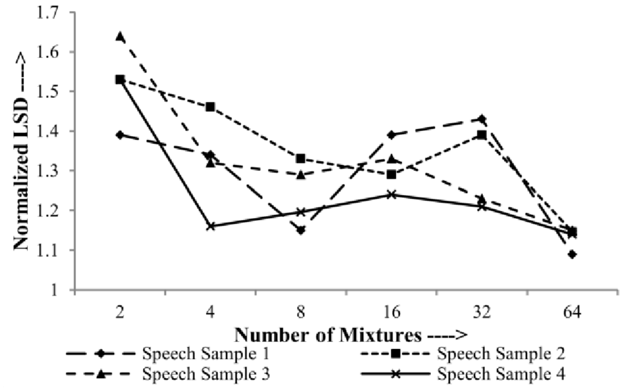


Fig. 4. Performance evaluation of GMM based model for different number of mixtures

VI. EXPERIMENTAL RESULTS

This section describes experimental set up for the performance evaluation of the proposed voice conversion system.

A. Database

In order to appraise the performance of the proposed algorithm ARCTIC database is used. Each speaker is recorded a set of 1132 phonetically balanced parallel utterances recorded at 16 kHz sampling frequency. This corpus include two female i.e. 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) [19]. The performance of proposed system is evaluated using subjective and objective measures. In this paper Mel Cepstrum Distortion (MCD) between target and synthesized speech is used as an objective measure to evaluate the performance of the system [12]. In order to evaluate the voice individuality and quality of the synthesized speech, subjective listening test are considered.

B. Experimental Conditions

TABLE I
SUMMARY OF THE EXPERIMENTAL CONDITIONS

Number of training utterances	50
Number of testing utterances	10
Type of analysis window	Hamming
Window length	30 milliseconds
Percentage Overlap	50 % (15 milliseconds)
Order of FFT	512
Number of Low time lifter samples	16
Number of High time lifter samples	464
RBF Spread Factor	0.01
Number of Mixture Models	64

C. Performance Evaluation

The performance of the proposed algorithm can be evaluated using subjective and objective techniques. In this paper, MCD between target and converted based objective measure is considered; it calculates the distortion between target and converted speech. It is highly correlated to subjective evaluation. Formant distortion is also considered as an objective measure to calculate the spectral distortion between converted and target speech. In order to evaluate the voice individuality and quality of the synthesized speech, subjective listening test are considered.

1) MCD based Evaluation

MCD between converted and target speech is calculate as [12],

$$MCD[dB] = \frac{10}{log_{10}} \sqrt{\sum_{i=1}^{D} mct^{t}{}_{i} - mcs^{c}{}_{i}} \qquad (21)$$

where, $mct^{t}{}_{i}$ and $mcs^{c}{}_{i}$ are the Mel Cepstrum Coefficients (MCC) of the target and synthesized speech respectively. The zeroth MCC term not considered in (14) since it describes the energy of the frame and is usually copied from the source. The performance of RBF and GMM based system is experimentally tested for different (10 to 40) number of source and target speaker of male and female from ARCTIC database. Results for the different source speaker to target speaker mapping are shown in figure 5. From results it can be observed that proposed voice conversion system performs better for intra-gender voice conversion than inter-gender voice conversion. It is also observed that cepstrum based features show better performance for RBF in comparison with GMM based model.

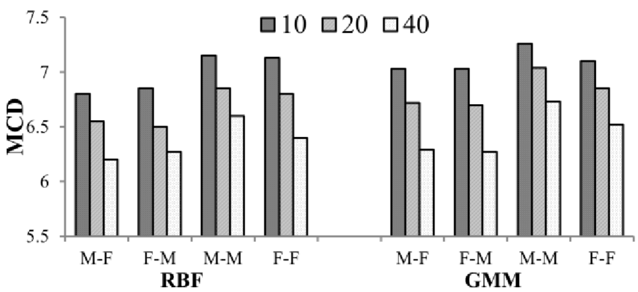


Fig.5. MCD based comparative performance of ANN and GMM Systems

2) Subjective Evaluation

The effectiveness of the algorithm can be evaluated using listening tests. These subjective tests are used to determine the closeness between the transformed and target speech sample. Inter gender and intra gender conversions are performed for which the sentences uttered are taken from the publically available ARCTIC database. The system is trained for 10-40 samples for the conversion from male-to-male, male-to-female, female-to-female and female-to-male. Ten synthesized speech utterances for each of the above mentioned cases and the corresponding target utterances were presented to twelve listeners. They were asked to judge their comparative performance with corresponding source and target 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 between the original target utterance and the transformed utterance and the other ratings indicate different levels of variation between 1 and 5. The ratings given to each set of utterances were used to calculate the mean opinion scores (MOS) for the four cases and the results are shown in figure 6. The obtained MOS results show that the conversion was effective, if the source and target were from different genders, i.e. the conversion was more efficient for male-to-female and female-to-male conversion as compared to male-to-male and female-to-female. The dissimilarity in the length of the vocal tract and the intonation patterns of different genders is the key 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 was also performed using the same set of utterances and speakers. In the ABX test, the listeners were 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 the value of ABX percentage, the more is the nearness of the transformed speech to the target utterance. The results of the ABX test are shown in figure 6.

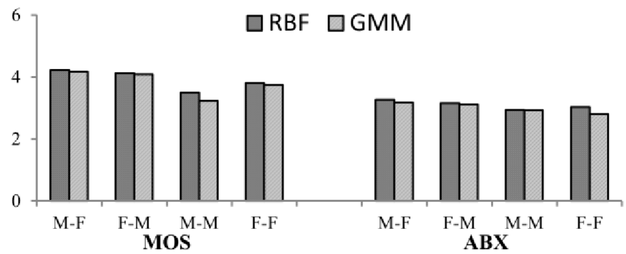


Fig. 6. Subjective evaluation performance of ANN and GMM Systems

VII. Conclusion

Using low time and high time liftering of the cepstrum, the vocal tract and glottal excitation parameters of the speech signal are separated. For capturing non-linear relationship between vocal tract and glottal excitation of speech frame RBF and GMM based models are developed. Comparative performance of the RBF and GMM based models are studied using subjective and objective evaluation techniques. The evaluation results indicate that RBF can be used as an alternative to the GMM based transformation model. The objective evaluation also verified that synthesized speech by both the techniques possess the characteristics of target speaker. The subjective evaluation convinced that the quality and naturalness of the transformed speech can be achieved with the proposed algorithm.

References

[1] L.M. Arslan, "Speaker transformation algorithm using segmental code books (STASC)" in *Speech Communication* (1999). 28 (3), pp. 211–226. Available: http://www.sciencedirect.com/science/article/pii/ S01676393 99000151.

[2] M. W. Kain, A. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction" in *IEEE Proc*. International Conference of Acoustics, Speech, Signal Processing (2001), vol.2, pp. 813–816. Available: http://ieeexplore. ieee.org.

[3] K. Lee, "Statistical approach for voice personality transformation" in *IEEE Trans*. on Audio, Speech, and Language Processing (2007). 15, pp. 641–651 Available: http://ieeexplore.ieee.org.

[4] C. Orphanidou, I. M. Moroz, S.J. Roberts, "Wavelet-based voice morphing" in *WSEAS Journal of Systems* (2004). 10 (3), pp. 3297–3302.Available: ww.robots.ox.ac.uk/~sjrob/Pubs/WSEAS_orphanid.pdf.

[5] J.H.Nirmal, S. S. Patnaik, M. A. Zaveri, "Voice Transformation using Radial Basis Function" in *Third International Conference* on Recent

Trends in Information, Telecommunication and Computing ITC 2012. Springer-Verlag Berlin Heidelberg, pp. 271-276.
[6] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks" in Speech *Communication* (1995). 16, pp. 207–216. Available: ww.sciencedirect.com/science/article/pii/016763939400058I
[7] H. Kuwabura, Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion" in Speech *Communication* (1995). 16, pp.165-173.
[8] M. Abe, S. Nakanura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization" in *IEEE Proc. of Int. Conf.* on Acoustics, Speech and Signal Processing (1988 May). pp. 655–658. Available: ik.itba.edu.ar/~pvacatel/files/papers/00196671.pdf.
[9] W. Verhelst, J. Mertens, "Voice conversion using partitions of spectral feature space" in *IEEE Proc.* of Int. Conf. on Acoustics, Speech, and Signal Processing (1996). pp. 365–368. Available: ieeexplore.ieee.org.
[10] N. Iwahashi, Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks" in *Speech Communication* (1995). 16, pp. 139–151. Available: www.sciencedirect.com/science/article/pii/ 016763939400051B.
[11] K . S . Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach" in *Computer Speech and Language Processing* (2010 July). 24, pp. 474–494. Available:www.sciencedirect. com/ science/article/pii/S0885230809000199.
[12] S. Desai, A.W. Black, B. Yegnanarayana, K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion" in *IEEE Trans.* Audio, Speech and Language Processing (2010 July). 18 (5) , pp. 954–964. Available: ravi.iiit.ac.in/publications/2010_Journ_P001.pdf.
[13] Stylianou, Y., Cappe, Y., Moulines, "Continuous probabilistic transform for voice conversion" in *IEEE Trans.* Speech and Audio Processing (1998). pp. 131–142. Available: www.ics.forth.gr/netlab/data/J1.pdf.
[14] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian Mixture Model with Dynamic Frequency Warping of straight spectrum" in IEEE *Proc.* International Conference of Acoustics, Speech, Signal Processing (2001). pp. 841–844. Available: http:\\ ieeexplore.ieee.org/iel5/7486/20369/00941046.pdf.
[15] T. Masuko, K. Tokuda, T. Kobayashi, S. Imai, " Speech synthesis using HMMS with dynamic features" in *IEEE Proc.* International Conference Acoustics, Speech, Signal Processing (1996). pp. 389–392. Available: crow.ee.washington.edu/people/bulyko/papers/HMM_Tokuda96.pdf.
[16] E.Helander,T.Virtanen,Jani N and M. Gabbouj, "Voice Conversion Using Partial Least Squares Regression" in *IEEE Trans.* On Audio, Speech, Language Processing (2010 July). 18(5), pp. 912-921. Available: www.cs.tut.fi/ sgn /arg/music/tuomasv/voco_pls.pdf.
[17] Alan V. Opphenheim , "Speech Analysis and Synthesis System based on Homomorphic filtering" in *Journal* of the Acoustical society of America. 45(2), pp. 458-465. Available: www.rle.mit.edu/dspg/ documents/ speechanaly_1969.pdf.
[18] Rodrigo C. Guidoa, Lucimar Sasso Vieiraa, Sylvio Barbon Juniora, "Neural Wavelet Architectures for voice Conversion" in *Science Direct Neurocomputing* (2007).71, pp. 174-180. Available: www.sciencedirect. com/science/article/pii/S0925231207002299.
[19] J. Kominek, A. W. Black, " CMU ARCTIC speech databases" in *Proceedings* 5th ISCA Speech Synthesis Workshop, Pittsburgh (2004 June),pp.223–224. Available: www.cs.cmu.edu/~awb/papers/ssw5/ arctic.pdf.