Applied Soft Computing 24 (2014) 1–12

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Voice conversion using General Regression Neural Network

Jagannath Nirmal[a,*], Mukesh Zaveri[b], Suprava Patnaik[c], Pramod Kachare[c]

[a] Department of Electronics Engineering, KJSCOE, Mumbai, India
[b] Department of Computer Engineering, SVNIT, Surat, India
[c] Department of Electronics Engineering, SVNIT, Surat, India

ARTICLE INFO

Article history:
Received 23 June 2013
Received in revised form 15 May 2014
Accepted 23 June 2014
Available online 11 July 2014

Keywords:
Gaussian Mixture Model
General Regression Neural Network
Pitch contour
Radial Basis Function
Voice conversion
Wavelet transform

ABSTRACT

The objective of voice conversion system is to formulate the mapping function which can transform the source speaker characteristics to that of the target speaker. In this paper, we propose the General Regression Neural Network (GRNN) based model for voice conversion. It is a single pass learning network that makes the training procedure fast and comparatively less time consuming. The proposed system uses the shape of the vocal tract, the shape of the glottal pulse (excitation signal) and long term prosodic features to carry out the voice conversion task. In this paper, the shape of the vocal tract and the shape of source excitation of a particular speaker are represented using Line Spectral Frequencies (LSFs) and Linear Prediction (LP) residual respectively. GRNN is used to obtain the mapping function between the source and target speakers. The direct transformation of the time domain residual using Artificial Neural Network (ANN) causes phase change and generates artifacts in consecutive frames. In order to alleviate it, wavelet packet decomposed coefficients are used to characterize the excitation of the speech signal. The long term prosodic parameters namely, pitch contour (intonation) and the energy profile of the test signal are also modified in relation to that of the target (desired) speaker using the baseline method. The relative performances of the proposed model are compared to voice conversion system based on the state of the art RBF and GMM models using objective and subjective evaluation measures. The evaluation measures show that the proposed GRNN based voice conversion system performs slightly better than the state of the art models.

1. Introduction

During the past two decades various voice conversion systems have been proposed to modify the speaker dependent parameters of the source speaker utterance so that it emulates those of the target (desired) speaker utterance [1,2]. Voice conversion is an emerging technology in speech processing used for commercial applications like the personification of text to speech, design of multi speaker based speech synthesis, security related applications, broadcasting and multimedia applications. In the film industry, it is used for voice editing, voice dubbing and voice animation [3,4]. Voice conversion could prove to be a simple and efficient way to create the desired variety of speakers without the need to record different speakers. In the field of medicine, voice conversion may improve the quality and intelligibility of laryngectomees voice which might repair there past speaking capabilities to produce natural voice [5].

Voice conversion is usually carried out in two different steps, namely training step followed by transformation step. In the training step, a set of speaker dependent features of the source and target speakers are derived and appropriate conversion function is established to map the source feature set onto that of the desired speaker's feature set. In the transformation step, the test speaker feature vector is modified using the conversion function derived in the training phase and the transformed speech signal holds the desired speaker characteristics [6]. In the training phase, the feature extraction is carried out by using speaker specific characteristics such as the shape of the vocal tract, the shape of source excitation and long term prosodic parameters [7]. Among these, the shape of the vocal tract and the shape of excitation parameters uniquely represents the speaker identity [7]. The vocal tract transfer function can be characterized using various speech features which can be classified into three categories: (1) features such as formant frequencies [8], which belong to acoustic phonetic model, (2) features derived, without speech production model such as, Cepstral Coefficients [9], Spectral Lines [3], and Mel-Cepstral Frequencies [9] and (3) Linear Predictive (LP) related features such as Linear Predictive Coefficients (LPC) [10], Line Spectral Frequencies (LSFs)

[*] Corresponding author. Tel.: +91 9167130111.
E-mail address: jhnirmal@somaiya.edu (J. Nirmal).

http://dx.doi.org/10.1016/j.asoc.2014.06.040
1568-4946/© 2014 Elsevier B.V. All rights reserved.

[11–13]. Amongst these feature representations, LSF overcomes the limitations of the LPC and results in much improved speech quality than any other features [13].

Various speaker specific models have been proposed in the literature to deal with the vocal tract mapping issue. Among these, the Vector Quantization (VQ) based codebook mapping [10] and Gaussian Mixture Model (GMM) are the most commonly used models [1,14]. In VQ, the source and target speakers utterances are clustered and a mapping rule for each cluster is formulated using minimum mean square error criteria. The main problem of VQ model is hard partitioning of the acoustic space produces discontinuities in the transition region. This affects the quality and naturalness of reconstructed speech signal [10,14]. An attempt [15] based on Fuzzy vector quantization and a Speaker Transformation Algorithm using Segmental Codebook (STASC) is further made to overcome the limitations of voice conversion system based on [2]. Dynamic Frequency Warping (DFW) [16] based voice conversion model is shown to yield the better quality of the converted speech signal. But, it transforms the formants to different frequencies without modifying the complete spectral shape leading to poor quality results. Therefore, DFW is further upgraded with weighted frequency warping technique [17].

Another approach based on GMM model uses the joint distribution of speech signal features. It partitions the speaker spectral space into overlapping classes. Then, a continuous probabilistic linear transformation function is defined from these partitions for parametric vector representation of the envelope [1,14]. But the quality and naturalness of the converted speech are found to be inadequate due to two reasons: First is the use of a large number of parameters in speech synthesis technique, and the second is over smoothing these parameters [1,14,18]. Therefore, various methods have been proposed to deal with the reconstruction issue of the GMM based model. It includes the use of precise conversion of phase spectrum, the use of robust vocoding methods such as STRAIGHT and the use of appropriate source filter models [8]. The over-smoothing issue is resolved via maximum likelihood estimators and hybrid methods [16,19,20,17]. Partial Least Square (PLS) regression based kernel transformation technique is also proposed [21] to capture the nonlinearities in the data to overcome the over-smoothing issues of GMM.

The speech synthesis technique based on Hidden Markov Model (HMM) produces the parameter vector when a new test input is given to a trained HMM model [22,23]. The resultant vector of the parameters is then used to synthesize the source speaker utterance by adopting HMM itself to the target speaker utterances [24,25]. However, the factors such as low quality of synthesized speech signal and over smoothing limits the usefulness of this approach. Followed by this nonlinear mapping capability of ANN is utilized to model dynamic pattern of vocal tract acoustic cues which itself is nonlinear in nature [11,12,9,13,26].

Along with the vocal tract parameters, the source excitation signal is also an important speech parameter which contains speaker individuality [1,27,28] so the high quality voice conversion system needs to transform source excitation parameter with suitable mapping function. Some of the existing techniques in the literature are: linear transformation, unit selection [28], codebook copying [10], residual prediction [29] and time delay neural network [30]. However, one of the main issues of the ANN is that it can capture the speaker specific characteristics empirically at the epoch level. This problem is further compounded by the fact that both the vocal tract characteristics and residual signals are predictive in nature. Therefore the conventional approaches of residual modification fail to capture the correlation between the source and desired speaker. It produces artifacts and phase distortion in reconstructing speech [12]. In order to overcome these issues, a new wavelet based technique is proposed to characterize the pitch residuals. The pitch residual is decomposed using wavelet transform and GRNN based mapping function is developed to capture the relation between source and target speakers. The prosody modification such as speaking style and short time energy of the speaker is highly desired for procurement of the converted speech utterances which are perceptually closer to the desired speaker utterances [1,6]. In the literature, various techniques like segment and sentence contour codebook, scatter plots, GMM based and linear models are experimented for pitch contour mapping [31,32,26,33]. In our work, the pitch contour of test speaker is modified according to the desired speaker using PSOLA as a baseline method [34]. Fixed scale factor derived from test and target speaker utterances is used to modify the energy profile of the test speaker according to that of the target speaker.

The physiological speech production structures for different speakers are highly nonlinear. The ANN and GMM based mapping functions are widely used to capture these nonlinearities in pattern. In spectral transformation approach using ANN and GMM models needs around 30–50 parallel utterances to formulate the voice conversion model. In addition both of these systems need to tune according to the amount of training data. An approach proposed in [9] has used Back propagation (BP) neural network to model the complex relations between the input–output feature vectors. The comparative analysis of spectral transformation between ANN and GMM based approaches using ARCTIC database is carried out in this approach. The results reveal the better performance of the ANN than that of the GMM based voice conversion model. Although this approach [9] is good in capturing the voice individuality and quality of the transformed speech, it also has some drawbacks. For instance, the performance of the network learning is strictly dependent on the shape of the error surface and connection weight initialization. The convergence to the global optimum is not guaranteed. Also the network architecture and parameters such as convergence rate and momentum needs to be determined by the user or by means of the optimized search. Besides the network parameters cannot be derived directly from the training examples. Therefore the selection of network parameters decides the success of network training for getting the required task done. Thus the time-consuming and iterative training procedure used in this approach limits its speed of convergence. Its another issue is only the convergence to local minima is guaranteed rather than global minima.

In this paper, we present a powerful method for voice conversion based on GRNN. It is a kind of Radial Basis Function (RBF) networks, developed by Specht. Our approach can basically speed up the learning procedure by utilizing the well known characteristics of the GRNN such as parallel architecture and single pass learning. This approach also guarantees the network convergence to the optimal regression surface when the number of samples becomes very large without necessitating any type of iterative training [35–38].

This paper mainly contributes to the following objectives: (1) Exploring GRNN based transformation model to capture the nonlinear mapping functions for modifying the LSF and wavelet decomposed LP-residual of a source speaker to that of a target speaker. (2) Evaluating the performance of GRNN based voice conversion system using subjective and objective methods. (3) Verifying that the proposed system performs better than that of the RBF and GMM based transformation models using evaluation measures.

The paper is organized as follows: the brief introduction of the present work. Section 2 provides the overview of the voice conversion system. Section 3 explains the state of the arts GMM, RBF transformation models used for voice conversion system. The proposed GRNN based transformation model is explained in Section 4. Section 5 explains the prosody modification in detail. Experimental results and analysis are described in Section 6. The overall conclusions of the paper are derived in Section 7.