Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Novel approach of MFCC based alignment and WD-residual modification for voice conversion using RBF

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

[a] Department of Electronics Engineering, K.J.Somaiya College of Engineering. Mumbai 400077, India
[b] Department of Computer Engineering, S.V.National Institute of Technology, Surat 395007, India
[c] Department of Electronics Engineering, S.V.National Institute of Technology, Surat 395007, India
[d] Department of Electronics Engineering, Veermata Jeejabai Institute of Technology, Mumbai 400031, India

ARTICLE INFO

Article history:
Received 7 March 2015
Received in revised form
30 May 2016
Accepted 31 July 2016
Communicated by R. Capobianco Guido

Keywords:
Dynamic time warping
Gaussian mixture model
LP-residual
Line spectral frequencies
Mel frequency cepstrum coefficient
Radial basis function
Residual selection method and
Wavelet packet transform

ABSTRACT

The voice conversion system modifies the speaker specific characteristics of the source speaker to that of the target speaker, so it perceives like target speaker. The speaker specific characteristics of the speech signal are reflected at different levels such as the shape of the vocal tract, shape of the glottal excitation and long term prosody. The shape of the vocal tract is represented by Line Spectral Frequency (LSF) and the shape of glottal excitation by Linear Predictive (LP) residuals. In this paper, the fourth level wavelet packet transform is applied to LP-residual to generate the sixteen sub-bands. This approach not only reduces the computational complexity but also presents a genuine transformation model over state of the art statistical prediction methods. In voice conversion, the alignment is an essential process which aligns the features of the source and target speakers. In this paper, the Mel Frequency Cepstrum Coefficients (MFCC) based warping path is proposed to align the LSF and LP-residual sub-bands using proposed constant source and constant target alignment. The conventional alignment technique is compared with two proposed approaches namely, constant source and constant target. Analysis shows that, constant source alignment using MFCC warping path performs slightly better than the constant target alignment and the state-of-the-art alignment approach. Generalized mapping models are developed for each sub-band using Radial Basis Function neural network (RBF) and are compared with Gaussian Mixture mapping model (GMM) and residual selection approach. Various subjective and objective evaluation measures indicate significant performance of RBF based residual mapping approach over the state-of-the-art approaches.

1. Introduction

The voice conversion system aims to adapt the acoustic characteristics of a given (i.e. source) speaker to a particular (i.e. target) speaker [1]. It employs two common stages: (i) training and (ii) transformation. In the training phase, voice conversion system identifies and extracts the speaker specific features from the utterances of both the source and the target speaker. These source and target features are employed to formulate the mapping function for capturing the nonlinear relations between speaker specific features. Afterwards, the transformation phase employs the trained mapping function to modify the features of the source speaker so as to make it perceptually similar to that of a target speaker [2–4]. The training phase of voice conversion involves acoustic modelling, feature alignment and acoustic mapping. The acoustic modelling signifies the shape of the vocal tract, shape of the glottal excitation and long term prosodic parameters [5–7]. Among these, the vocal tract parameters are relatively more prominent for identifying the speaker uniqueness than the source excitation parameters [6,8].

Various methods for feature extraction have been proposed in the literature to characterize the vocal tract parameters of the speech frame, namely, formant frequency [5], formant bandwidth [5,9], Linear Prediction Coding (LPC) [10], cepstrum coefficient [11], Mel Cepstrum Envelope (MCEP) [12], Mel Generated Cepstrum (MGC) [1] and Line Spectral Frequencies (LSFs) [13–15]. Amongst these feature representations, LSF results in much more improved speech quality than any other features [15]. The glottal excitation signal is another important parameter conveying the essential information about speaker identity [8].

In high quality voice conversion system, the alignment of the source and target samples is of utmost importance to have parallel data prior to the estimation of mapping functions. Time

* Corresponding author.
E-mail addresses: jhnirmal@somaiya.edu (J. Nirmal), mazaveri@gmail.com (M. Zaveri), suprava_patnaik@yahoo.com (S. Patnaik), pramod_1991@yahoo.com (P. Kachare).

http://dx.doi.org/10.1016/j.neucom.2016.07.048
0925-2312/

consuming manual alignment technique can be easily traded by employing an automatic time alignment technique called as a Dynamic Time Warping (DTW) [16]. Although, the conventional voice conversion systems present number of different alignment procedures. But to our knowledge a comparative study of different issues involved in appropriate alignment is not available.

Our present work deals with some of the limitations of conventional LSF based warping method and provide a new alignment method using MFCC based warping path, which improves the conversion performance of the proposed system. It is shown through experimental analysis that the novel approach based on MFCC features gives better results in terms of speech quality than the conventional LSF-DTW technique [17]. The final step of training is to obtain the mapping function. Several speaker specific models have been proposed in the literature to deal with the vocal tract mapping [1,2,8,10,12,13,16]. However, very few studies have been carried out concerned with residual signal transformation models. These methods can be categorized as residual copying [18], residual prediction [19], residual selection [20] and unit selection [14]. The residual copying and residual prediction methods face the problem of losing the strong correlation between the source and system characteristics [15]. The residual selection method involves high computational cost, so it limits the transformation ability to codebook size [20].

Our present work deals with the high dimensionality issue of residual signal so as to decrease the time consuming computations and the complexity of the transformation model [21]. This approach also tackles the issue of artifacts generated in consecutive frames. Additionally, the voice conversion system may be improved by using actual transformation techniques for LP residual signal instead of using conventional statistical selection methods. Therefore, in the proposed approach a RBF based transformation model is used. We have derived optimum RBF mapping functions to map the LSF, filter gain and LP-residual features. The conversion quality of the LP residual sub-band based proposed approach and the state of the art residual selection approach is examined through various performance measures.

The contents of this paper are structured as follows: Section 2 explains the proposed speech alignment mechanism. Wavelet sub-band processing of residual signal is described in Section 3. Section 4 gives details of the proposed voice conversion framework. The RBF and GMM are explored to derive the mapping functions for modifying the vocal tract and the glottal excitation in Section 5. The baseline residual selection approach is presented in Section 6. The experimental results and comparative evaluations including objective and subjective measures are specified in Section 7. Finally the conclusion is discussed in Section 8.

2. Alignment of parallel data

The proposed voice conversion system requires parallel training data from both the source and the target speakers, usually DTW is used for frame level alignment between the speech signals of the source and the target speakers [16]. In well studied voice conversion systems, DTW is used for alignment of LSF based features and generates pseudo-aligned data. Although, the standard linear prediction features provide information about the formants (i.e. spectral peaks), but ignores the valleys (i.e.spectral zeros) of the spectrum completely [1]. Therefore, the conventional LSF based DTW alignment approach fails to capture nasal, plosives and unvoiced parts of the speech signal. These essential parts of the speech signal are recognized by the valleys in the spectrum [17]. In order to resolve these issues, we have proposed MFCC based approach for alignment, which provides more reliable spectrum in terms of peaks as well as valleys. As the MFCC approach gives

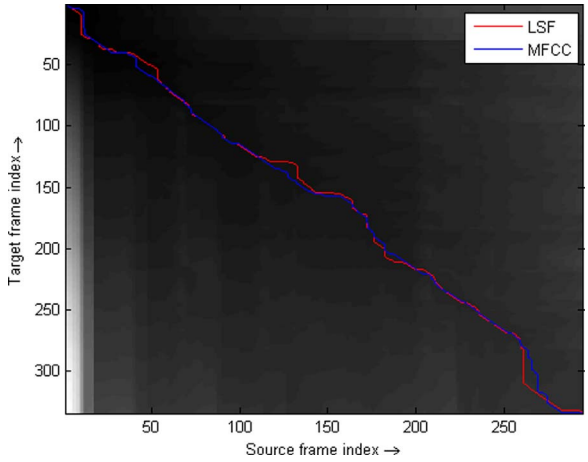

Fig. 1. Dynamic time warping paths for LSF and MFCC based feature vectors.

information about voiced as well as unvoiced articulations [1], it becomes a more suitable alternative for measuring the acoustic distances than that of the DTW approach.

The spectral characteristics of LSF and MFCC based warping paths for single source-target feature vectors are illustrated in Fig. 1. The warping paths presented here can be analysed in three different parts as shown in Fig 2.:

(a) *Diagonal segment*: It corresponds one-to-one spectral matching between source and target feature vectors (i.e. frames). It causes matching of mutually exclusive source–target pair over the complete feature set.

(b) *Horizontal segment*: This can be understood as many-to-one and spectral matching between multiple source feature vectors and isolated target feature vector. This type of warping path suggests repetitions of target feature vectors (i.e. frames) in time aligned sample.

(c) *Vertical segment*: Apparently, we have third part in warping path as one-to-many and which minimizes the spectral distance for multiple target feature vectors for unique source feature vector. Like prior case, warping path suggests repetitions of source feature vectors (i.e. frames) in time aligned sample.

Mathematically, a complete warping path (W) can be formulated as a combination of bijective, surjective and injective mapping functions over source (Fs) and target (Ft) frame space,

$$W = \begin{Bmatrix} f_s \rightarrow f_t & f_s \epsilon F_s, f_t \epsilon F_t \\ F_s \rightarrow f_t & f_t \epsilon F_t, F_s \subset F_s \\ f_s \rightarrow f_t & f_s \epsilon F_s, F_t \subset F_t \end{Bmatrix} \quad (1)$$

where (F_s) and (F_t) are source and target frame space.

The conventional alignment technique consists of all the above three possible types of segments causing unwanted frame repetitions on both source and target side. These repetitions in the aligned feature set may result in overtraining issue, which in turn limits the efficiency of transformation model. The repetitions in aligned feature set can be lessened by considering either of the source or the target so that unique feature set is derived.

The constant source alignment is a many-to-one technique, which uniquely maps each source feature vector to appropriate target feature vector based on minimum Euclidean distance. Similarly, the constant target alignment is one-to-many technique, which uniquely maps each target feature vector to corresponding source feature vector. The many-to-one mapping causes some of the frames in target to be repeated while few of them to be excluded in the aligned target feature set and vice a versa. In constant source alignment, the unique source feature vectors remove all the vertical segments. It leaves only a combination of diagonal

Please cite this article as: J. Nirmal, et al., Novel approach of MFCC based alignment and WD-residual modification for voice conversion using RBF, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.07.048