Available online at www.sciencedirect.com

ScienceDirect

NEUROCOMPUTING

Neurocomputing 71 (2007) 174–180

www.elsevier.com/locate/neucom

A neural-wavelet architecture for voice conversion

Rodrigo Capobianco Guido[a,*], Lucimar Sasso Vieira[a], Sylvio Barbon Júnior[a], Fabrício Lopes Sanchez[b], Carlos Dias Maciel[b], Everthon Silva Fonseca[b], José Carlos Pereira[b]

[a] *SpeechLab/FFI/IFSC/USP—Department of Physics and Informatics, Institute of Physics at São Carlos,University of São Paulo, Avenida Trabalhador São Carlense 400, 13560-970 São Carlos, SP, Brazil*

[b] *SEL/EESC/USP—Department of Electrical Engineering, School of Engineering at São Carlos, University of São Paulo, Avenida Trabalhador São Carlense 400, 13560-970 São Carlos, SP, Brazil*

Available online 29 August 2007

Abstract

In this letter we propose a new architecture for voice conversion that is based on a joint neural-wavelet approach. We also examine the characteristics of many wavelet families and determine the one that best matches the requirements of the proposed system. The conclusions presented in theory are confirmed in practice with utterances extracted from TIMIT speech corpus.

Keywords: RBF neural networks; Wavelet transforms; Voice conversion

1. Introduction

Voice conversion, also known as voice morphing, enables a source speaker to transform his speech pattern to sound as if it were spoken by another person, that is the target speaker, preserving the original content of the spoken message [18].

Much literature has recently appeared addressing the issue of voice conversion [12,5,17]. Most methods are single-scale methods based on the interpolation of speech parameters and modeling of the speech signals using formant frequencies (FF) [1], linear prediction coding (LPC) and cepstrum coefficients (CC) [6], line spectral frequencies (LSF) [11], segmental codebooks (SC) [16], besides many others. We can also find some techniques that are based on hidden Markov models (HMMs) and Gaussian mixture models (GMM) [4], that are well-known methods in the speech community. Most of the techniques we mentioned suffer from absence of detailed information during the extraction of formant coefficients and the excitation signal. This results in the limitation of being able to accurately estimate parameters as well as distortion caused during the synthesis of the target speech. Turk and Arslan [16] introduced the discrete wavelet transform (DWT) [2,14] for voice conversion and got encouraging results. Following the ideas of Turk and Arslan, we can find other interesting contributions, as the one of Orphanidou et al. [13].

Particularly, this paper proposes a new algorithm for voice conversion that is based on wavelet transforms and radial basis function (RBF) neural networks [8]. This is the main contribution of this work, that extends the considerations of Turk and Arslan, and Orphanidou et al., on the use of wavelets for voice conversion.

This paper is organized as follows. Section 2 presents a brief overview on how the wavelet-based algorithms for voice conversion work. The proposed approach is presented in Section 3. A study on wavelets and their important characteristics for voice conversion, in order to determine the best wavelet family to be used, is available in Section 4. Section 5 describes the tests and results, and, lastly, Section 6 presents the conclusions.

2. A brief review on wavelet-based voice conversion

The basic idea behind the use of the DWT for voice conversion is the sub-band separation. With that, the pitch period of voiced sounds, FF [4], plus other information,

[*] Corresponding author. Tel./fax: + 55 16 33739777.
E-mail address: guido@ifsc.usp.br (R.C. Guido).
URL: http://speechlab.ifsc.usp.br (R.C. Guido).

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.08.010

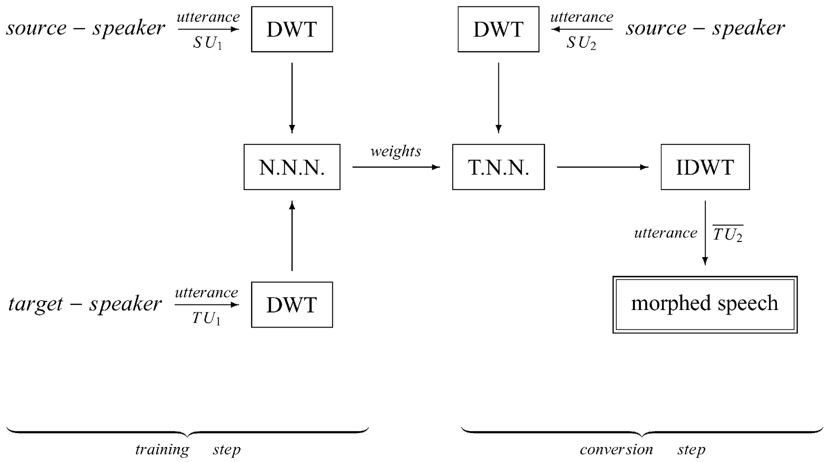

Fig. 1. Basic architecture of a wavelet-based system for voice conversion. (N.N.N.): neural network not yet trained; (T.N.N.): trained neural network.

can be effectively treated and converted into separate manners. In order to convert the source speaker's pattern into the target speaker's pattern, artificial neural networks can be used, as in [13]. Usually, these networks, that are important components of the system, are multilayer perceptron or RBF networks [8].

For training the networks, the DWT of the source and target speakers' voice signals are used as input. The sentences or phonemes uttered by both speakers must be the same. The weights of the hidden and output layers of the networks are adjusted in such a way that the source speaker's pattern is converted into the target speaker's pattern. This conversion is done for each sub-band of frequencies separately, and the training step is a supervised learning procedure [8].

After the training step, the conversion step receives as input a different sentence uttered by the source speaker, applies the DWT, and sends the signal to the networks that convert it. Lastly, the inverse DWT (IDWT) is applied to return the converted signal to the time domain. The converted sentence sounds as if it were spoken by the target speaker. Fig. 1 illustrates this.

Additional details about the architecture of the neural networks that can be used for voice conversion were not included in this section because Orphanidou et al discuss this topic with details in [13], furthermore, complementarily explanations are given in Section 5 of this letter.

3. The proposed approach

The proposed approach is divided into two parts: training (TR) and testing (TE). They are completely described in Tables 1 and 2, respectively, and further explanations follow.

All the RBFs use the Gaussian's equation [13] as activation function in the hidden layers. On the other hand, the output neurons use a simple linear weighted function. To train the RBFs, a two-step procedure was adopted. In the first step, a non-supervised training is used to determine the centers and variances of the Gaussian Kernels, i.e., only the source's speech data is used. Then, the weights of the output layer are adjusted by means of a supervised procedure, i.e., the source and target's speech data are used. This is exactly the procedure adopted by Orphanidou et al. in [13].

In our approach, each one of the 21 RBFs, R_i, $(0\leqslant i\leqslant 20)$, has a_i input and output neurons, and b_i hidden neurons; a_i being the number of samples of the critical band i within a speech frame of 256 samples, according to Table 3, and b_i being the number of speech frames to be used during the training, that is the same for all RBFs. This allows the supervised part of the training to be easily solved by using the Moore–Penrose's pseudo inversion in singular values [9].

Another important observation is that only the voiced speech frames are used during the training, since unvoiced data does not contain significant information of the vocal tracts' resonances.

4. Exploring the characteristics of wavelets

According to the DWT theory [2], the jth-level decomposition of a given discrete (speech) signal, $f[n]$, can be written as [7]

$$f[n]=\sum_{k=0}^{n/2^j-1} R_{j,k}[n]\phi_{j,k}[n]+\sum_{t=1}^{j}\sum_{k=0}^{n/2^j-1} S_{t,k}[n]\psi_{t,k}[n], \qquad (1)$$

$\phi[n]=\sum_k h[k]\phi[2n-k]$ and $\psi[n]=\sum_k g[k]\phi[2n-k]$ being the scaling and wavelet functions, respectively, that form a Riesz basis [2] to write signal f, $R_{j,k}[n]\leqslant\langle f,\phi_{j,k}[n]\rangle$, $S_{t,k}[n]\leqslant\langle f,\psi_{t,k}[n]\rangle$, and $h[k]$ and $g[k]=(-1)^k h[N-k-1]$ being the quadrature mirror (QMF) low-pass and high-pass analysis filters [2], respectively.

During the training and conversion steps, when $f[n]$ is being *analysed*, as illustrated in Fig. 2, low-pass and high-pass filtering occur by discrete convolutions, followed by down-samplings by 2, and, therefore, the length of $h[k]$ and $g[k]$, N, is responsible for both *frequency selectivity*, Q, and *time resolution*, R. As N increases, Q increases and R