A NEW VOICE TRANSFORMATION METHOD BASED ON BOTH LINEAR AND NONLINEAR PREDICTION ANALYSIS

Ki Seung Lee, Dae Hee Youn, and Il Whan Cha

Center for Signal Processing Research
Dept. of Electronic Eng., Yonsei university
Seoul, 120-749, Korea.
E-mail:jlks@stellar.yonsei.ac.kr

ABSTRACT

In this paper, we describe a voice transformation meth-od which changes source speaker's acoustic features to those of a target speaker. The method developed here, acoustic features are divided into two parts, linear and nonlinear parts. Linear parts are characterized by LPC cepstrum coefficients which are obtained from LP analysis. As for nonlinear part, which represent the excitation signal, is modelled by the long-delay nonlinear predictor using a neural net. Conversion rules for excitation signal are generated by the average pitch ratio and the mapping codebook, and those for LPC cepstrum coefficients are based on the orthogonal vetctor space conversion. In addition, the spectral envelope compensation is proposed to correct spectral distortion in the transformed speech. A listening test shows that the proposed method makes it possible to convert speaker's individuality while maintaining high quality.

1. INTRODUCTION

Voice transformation is a process of changing voice personality; i.e. speech uttered by a source speaker is modified to sound as if a target speaker had uttered it. This technique has numerous applications that include personalification of text-to-speech synthesis system, preprocessing for speech recognition, improving the effectiveness of foreign language training system, and so on [5][8][9][10].

Voice transformation is performed in two steps. In training stage, acoustic parameters of the speech signals uttered by both source and target spearkers are computed and appropriate rules mapping the acoustic space of the source speaker onto that of the target speaker are obtained. In the transformation stage, the acoustic features of the source signal are transformed using the mapping rules such that the synthesized speech possesses the personalities of the target speaker.

It is well known that the vocal tract transfer function is the dominant factor in specifying speaker individuality [7]. For this reason, previous methods have mainly been dedicated to the transformation of vocal tract transfer function which is represented by the linear prediction coefficients [5][8][9][10]. We have proposed a conversion method for vocal tract transfer function which is based on vector space approach [5]. According to this approach, transformation is applied only to the principle components of speech signal. The performance of transformation was acceptable, while feaure vector undergoes a dimensionality reduction. However, there exist some spectral discrepancies between transformed signal and target signal due to the incomplete transformation of excitation signal. To solve this problem, a spectral envelope compensation method is proposed in this paper.

On the other hand, LP-residual is an important fator in preserving naturalness of transformed speech [7]. Because pitch complexes in the residual after LPC analysis of voiced speech are highly nonlinear [1], the transformation method based on linear model would yield a some "hollow" and "muffled" quality sound. To overcome this drawback, we use nonlinear predictor for modelling excitaion signal. Recent studies have shown that nonlinear prediction can be implemented with time-delay neutal net (TDNN) [2]. In this paper, we take advantage of the nonlinear prediction capability of TDNN and apply it to the developement of the excitaion signal transformation method.

The outline of the paper is as follows. Section 2 presents the methods of modelling and transformation of the excitation signal. In section 3, LPC cepstrum transformation method is presented. In section 4, experiments and the results of proposed method are described. Finally, conclusions are made in section 5.

2. MODELLING AND TRANSFORMATION OF EXCITATION SIGNAL

2.1. Modelling the excitation signal

The speech signal contains more or less non-linearities, as that are typically found in LP residual [1]. To cope with these properties, several nonlinear predictors have been proposed; such as 2nd order volterra filter [1][3], radial basis functions (RBF) [4] , and time delay neural net (TDNN) [1][2].

Among these, neural net-based predictors can be used for modelling data without any specific prior assumption about the form of nonlinearity [2]. Other advantages of neural net predictor are that the number of filter coefficients grows slowly with the prediction order, and the analysis filter and the synthesis filter are always stable [1].

We apply the neural net predictor to modelling the exci-

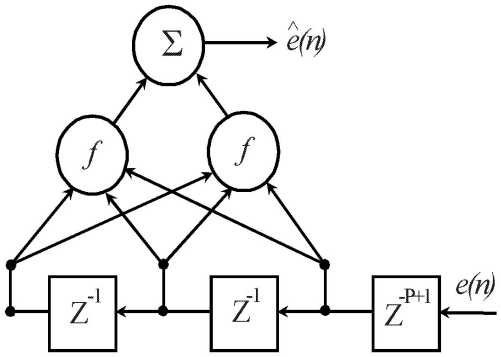

Figure 1: Long delay neural net predictor.

tation signal. A neural net model of a nonlinear long delay predictor is shown in Figure 1. It consists of three layers, which contain three, two and one units in the input, hidden and output layers, respectively. This predictor can be described by the following equations:

$$\begin{aligned}\hat{e}(n) &= \sum_{j=1}^{2} w_j f(\sum_{i=-1}^{1} v_{ji} e(n-P-i)) \\ &= F(\mathbf{W}, \mathbf{V}, \mathbf{e}(n)) \end{aligned} \quad (1)$$

where $\hat{e}(n)$ is predictive exctiation signal and w_j, v_{ji} are weight values for output and iternal hidden layer, respectively. In this paper, we define $f(.)$ as the commonly used sigmoid function, and P is the estimated pitch period for current frame. The weighs w_j, v_{ji} are trained to minimize the square of predictive residual, $||e(n) - \hat{e}(n)||^2$ using the backpropagation algorithm [2]. In the method proposed, a nonlinear predictive vector quantizer (NLPVQ) [2] is used to reduce the computational loads in parameter estimation process and represent the exctation signal with a limited number of parameters. This consists of a set of predictors $\{F(\mathbf{W}_k, \mathbf{V}_k, \mathbf{e}(n)), k = 1, ..., K\}$ which minimize the power of predictive residual. During quantization, each frame of excitation signal is successively applied to all the predictors in the VQ codebook. The predictor with the least predictive error is then selected to quantize the current frame. The total distortion is given by

$$D_{tot} = \sum_{n=1}^{N} argmin_k ||e(n) - F(\mathbf{W}_k, \mathbf{V}_k, \mathbf{e}(n))||^2 \quad (2)$$

In NLPVQ design process, LBG algorithm [6] which is widely used as VQ training method leads to poor performance. This is mainly because averaging and using the Euclidean distance measure is unsuitable for neural net.

To improve the vector quantizer design for neural net predictor, modification of obvious LBG algorithm are introduced. Instead of minimizing distortion between orignal vectors and coded vectors in LBG algorithm, we minimize the cost function be the sum of the residual energies, D_{tot} in (2). and that the updating equations becomes:

$$\Delta \mathbf{W}_k = \eta \sum_{\mathbf{e}(n) \in k} \sum_{n=1}^{N} \nabla_{\mathbf{w}_k} \frac{1}{2} ||e(n) - F(\mathbf{W}_k, \mathbf{V}_k, \mathbf{e}(n))||^2 \quad (3)$$

$$\Delta \mathbf{V}_k = \eta \sum_{\mathbf{e}(n) \in k} \sum_{n=1}^{N} \nabla_{\mathbf{v}_k} \frac{1}{2} ||e(n) - F(\mathbf{W}_k, \mathbf{V}_k, \mathbf{e}(n))||^2 \quad (4)$$

The performance of the proposed neural net predictor is summarized in table 1. The results in table 1 were calculated over a total of 1000 frames from one speaker. The prediction gain of neural net predictor with 64 predictors is 6.15 dB, which is 1.5 dB greater than that of 3 tap long-delay linear predictor. Inspecting the nonlinear predictive residual, it can be seen that residual contains a small amount of pitch pulses which are often found in the linear predictive residual. These results confirm that most information in the excitation signal can be represented by the nonlinear neural net predictor.

Table 1. Prediction gain of two predictors.

No. of predictor	Prediction Gain (dB)	
	Nonlinear	Linear
32	5.98	4.03
64	6.15	4.62

2.2. Transformation of the excitation signal

The transformation of excitation signal is accomplished by changing pitch period P in (1), and mapping the codebooks [8] of the two speaker. This process is depicted in Figure 2. The average pitch period of one speaker contribute a great deal to speech individuality [7]. Hence, source speaker's pitch period is modified by the average pitch modification factor β_{ave}; defined by

$$\beta_{ave} = \frac{\overline{P}_{source}}{\overline{P}_{target}} \quad (5)$$

where $\overline{P}_{source}$, $\overline{P}_{target}$ are the average pitch periods of the source and target, respectively.

The mapping codebook describes a mapping rule between two vector spaces that are obtained from source and target speaker's neural net nonlinear predictors, respectively. These are constructed by following training process. First, both source and target speakers pronounce the same training word set. For each word, the correspondence between vectors obtained from the two speakers is determined using Dynamic Time Warping (DTW). And the vector correspondences between two speakers are accumulated as histogram. Finally, the mapping codebook for nonlinear predictor is defined based on the maximum occurrence in the histogram. As a result, all vectors in the source codebook have one-to-one corresponding vectors in the target codebook. After determining the average pitch ratio and the mapping codebook, the transformed excitation signal is synthesize by the equation:

$$\begin{aligned}\hat{e}_t(n) &= \sum_{j=1}^{2} \hat{w}_{kj} f(\sum_{i=-1}^{1} \hat{v}_{kji} \hat{e}_t(n - \beta_{ave} P - i)) + r(n) \\ &= F(\hat{\mathbf{W}}_k, \hat{\mathbf{V}}_k, \hat{\mathbf{e}}_t(n)) + r(n) \end{aligned} \quad (6)$$

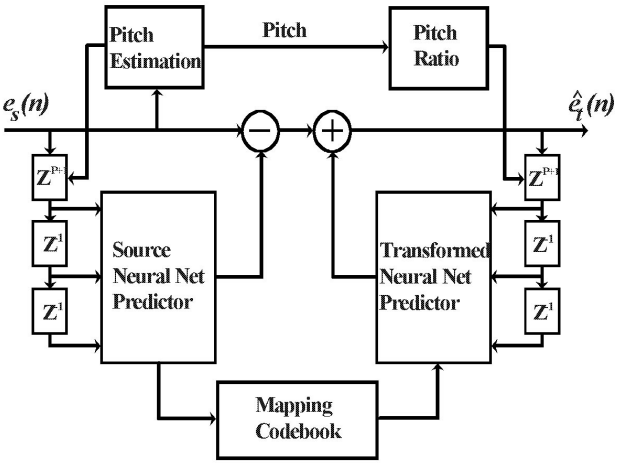

Figure 2: Blockdiagram of excitation transformation.

where $r(n)$ is predictive residual of source excitation signal given by

$$r(n) = e_s(n) - F(\mathbf{W}_k, \mathbf{V}_k, \mathbf{e}_s(n)) \tag{7}$$

and $\hat{\mathbf{W}}_k, \hat{\mathbf{V}}_k$ are the target vectors corresponding to the source vector $\mathbf{W}_k, \mathbf{V}_k$, respectively. Pitch peroid is meaningful in the voiced frames, transformation of the excitation signal is applied only to the voice part.

3. TRANSFORMATION OF LPC CEPSTRUM

The mapping rule of the LPC cepstrum coefficients is based on the orthogonal vector space conversion [5]. The underlying principle is to represent one speaker's LPC cepstrum vector as a signal vector in his(or her) own orthogonal vector space. The vector set of these space is composed of eigenvectors whose eigenvalues are greater than given threshold. Thus, LPC cepstrum vector can be represented by a reduced number of coefficients while preserving the detailed structure of the spectral envelope. Voice personality transformation is implemented by substituting the vector space of the source speech with that of the target speech, and moving all the vectors in the source vector space to desired points in the target vector space. According to this model, source and target LPC cepstrums, $\mathbf{C}_i^s$, $\mathbf{C}_i^t$ are expressed as a weighted sum of the principle eigenvectors, $\mathbf{e}_m^s$, $\mathbf{e}_m^t$.

$$\mathbf{C}_i^s = \sum_{m=1}^{M^s} s_m^i \mathbf{e}_m^s \quad , \quad \mathbf{C}_i^t = \sum_{m=1}^{M^t} t_m^i \mathbf{e}_m^t \tag{8}$$

where M^s and M^t are the number of principle vectors for source and target speeches, respectively. The transformed LPC cepstrum is given by

$$\hat{\mathbf{C}}_i^t = \sum_{m=1}^{M^t} \hat{t}_m^i \mathbf{e}_m^t \tag{9}$$

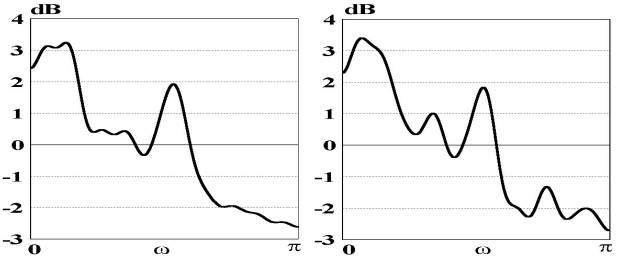

Figure 3: Spectral envelope, Left : from the transformed LPC cepstrum, Right : from the transformed speech signal.

where $\hat{t}_m^i$ is a tranformed weighting value, and is obtained from the equation

$$\hat{t}_m^i = \sum_{n=1}^{M^s} h_{mn} s_n^i + o_m \ , \quad m = 1, 2, ..., M^t \tag{10}$$

and h_{mn}, o_m are determined in training procedure, by minimizing MSE between $\hat{t}_m^i$ and t_m^i. Although source LPC cepstrum is transformed by (10), the spectral envelope of transformed speech is not excatly same as the spectral envelope obtained from the transformed LPC cepstrum. As shown in Figure 3, distinct difference is found in the high frequency region. This is mainly because the flatness of the transformed excitaion signal is not guaranteed. To correct this problem, a spectral envelope compensation is introduced in this paper. Defining the spectral envelopes from the transformed speech and transformed LPC cepstrum as $\hat{\mathbf{H}}_\mathbf{t}(\omega)$ and $\hat{\mathbf{H}}_\mathbf{c}(\omega)$, respectively, the compensated STFT of transformed speech is given by

$$\hat{\mathbf{S}}_\mathbf{c}(\omega) = \frac{\hat{\mathbf{H}}_\mathbf{t}(\omega)}{\hat{\mathbf{H}}_\mathbf{c}(\omega)} \hat{\mathbf{S}}_\mathbf{t}(\omega). \tag{11}$$

The final transformed speech is constructed by IDFT of $\hat{\mathbf{S}}_\mathbf{c}(\omega)$. As the higher order formant frequencies play an important role in interspeaker variability [7], the quality of speech signal obtained from the above method is more similar to that of the target speech.

4. EXPERIMENTS AND RESULTS

Experiments were performed to evaluate the performace of the proposed voice transformation method. The database used in order to train the mapping rules consists of the 61 words of the korean language uttered by two different male speakers. Speech signals were digitized at 10KHz sampling frequency and the order of LPC cepstrum was set to 20. Two evaluation tests for the proposed method were carried out, the capability of converting excitation signal, and the quality of transformed speech.

In figure 4, source, target, and transformed excitation signals are shown. Results in this figure shows the capability of proposed excitation transformation method. The most voiced parts of source speech signal exhibited this good result, but somewhat "buzzy" quality or click noise are noticeable in regions of speech which contain mixed voicing.

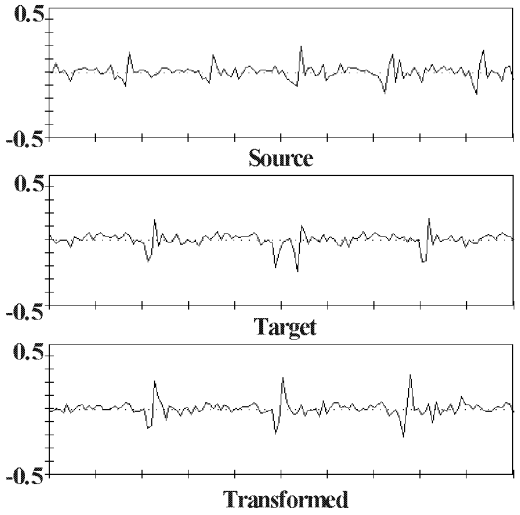

Figure 4: Excitation signals of source (upper), target (middle), and transformed (bottom). Both target and tranformed signals are time-aligned.

The major reason for this degradations is due to the incomplete pitch estimation and voice/unvoice decision. Thus, it should be mentioned that the performance improvement is possible by employing more robust pitch esitmation and voiced/unvoiced decision algorithms.

We presented 4 utterances to 10 listeners for a subjective listening test. The first two utterances were the original source and target signals. The third one is the tranformed speech obtained from excitation spectrum scaling method [9] and no spectrum compensation . And fourth one is generated by proposed method. These two speeches are obtained from the same vocal tract transformation method [5]. Subjective listening test was consisted in two categories. We checked out whether each listener correctly identifies the transformed speech as a target speechs and which utternce are more preferable to each listener. The preference test are performed in the region of voice part. The two methods exhibited the equal correct identification ratio. However, the preference test showed superior performance of the proposed method. Among the ten listeners, seven judged that the proposed method produce more natural-quality transformed speech. Another listeners said that both kinds of transformed speech were not significantly different. These results indicate that the effectiveness of the proposed transformation method.

5. CONCLUSION

We proposed a new voice transformation method based on the codebook mapping of nonlinear predictor and vector space conversion of LPC cepstrum coefficients. The proposed meth-od showed its ability in both changing speaker personality and preserving naturalness. In the method proposed, excitation signal is modelled by the nonlinear predictor using a long delay neural net. Although the excitation siganl contains less speaker's individuality than LPC cepstrum, manipulating the excitation signal is important factor to obtain natural quality sound. Experimental results were certified this fact. Thus, to obtain the transfomred speech sound as if a target speaker had really uttered, the excitaion signal should be modelled and transformed by more improved method. This work remains as a future study.

6. REFERENCES

[1] J. Thyssen, H. Nielsen, and S. D. Hansen, "Non-Linear Short-Term Prediction in Speech Coding," *Proc. ICASSP*, pp.I185-I188, 1994.

[2] L. Wu, M. Niranjan, and F. Fallside, "Nonlinear Predictive Vector Quantization with Recurrent Neural Nets," *Proc. IEEE-SP Workshop on Neural Networks for signal Processing*, pp. 372-381; Baltimore, MA, 1993.

[3] E. Mumolo, A. Carini, and D. Francescato, "ADPCM with Non Linear Predictors," *Proc. EUSIPCO-94*, vol. I, pp.387-390; Edinburgh, Scotland, U.K., 1994.

[4] F. Diaz-de-Maria, and A. R. Figueiras-Vidal, "Nonlinear Prediction for Speech Coding Using Radial Basis Functions," *Proc. ICASSP*, pp. 785-788, 1995.

[5] K. S. Lee, D. H. Youn, and I. H. Cha, "Voice personality transformation using an orthogonal vector space conversion," *Proc. EUROSPEECH-95*, vol. 1, pp.427-430; Madrid, Spain, 1995.

[6] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. on Communications*, vol. 28, Jan., 1980. pp.21-24, 1995.

[7] D. G. Childers, B. Yegnanarayana, and Ke Wu, "Voice Conversion: Factors Responsible for Quality," *Proc. ICASSP*, pp.748-751, 1985.

[8] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice Conversion through Vector Quantization," *Proc. ICASSP*, pp. 655-658, 1988.

[9] I. H. Nam, "Voice personality transformation," Ph. D. Thesis, Electrical Engineering, Rensselaer Polytechnic Institue, Troy, New York, 1991.

[10] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, pp. 175-187, 1992.