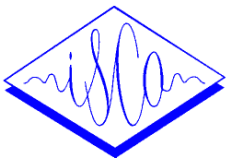
ISCA Archive
http://www.isca-speech.org/archive



4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

VOICE PERSONALITY TRANSFORMATION USING AN ORTHOGONAL VECTOR SPACE CONVERSION

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

*Department of Electronics Engineering
Yonsei University
Sudaemoon-Ku, Seoul, KOREA
e-mail : jlks@stellar.yonsei.ac.kr

10.21437/Eurospeech.1995-116

ABSTRACT

A newly developed voice personality transformation algorithm is introduced in this paper. Voice personality transformation is the process of changing one person's acoustic features (source) to those of another person (target). In this paper, personality transformation is achieved by changing the LPC cepstrum coefficients, excitation spectrum and global/local pitch contour. An orthogonal vector space conversion technique is proposed to transform LPC the cepstrum coefficients. This technique consists of principle component decomposition by applying the Karhunen-Loeve(KL) transformation and minimum mean-square error coordinate transformation(MSECT). To transform prosodic characteristics, we propose a simple pitch contour modification method. The experimental results show the effectiveness of the proposed algorithm in both subjective and objective evaluations.

1. INTRODUCTION

Voice personality transformation is the technique of making one person's voice (source) sound like another person's voice (target). This technique has many potential applications in the areas of speaker adaptation of automatic speech recognition systems and personalization of speech synthesis systems.

In voice personality transformation, there are several important problems to be considered. One of these is to find an optimal mapping rule which converts the acoustic feature parameters of the source signal to those of the target signal. Previous voice personality transformation methods used artificial neural networks[2], and vector quantization combined with a nearest neighbor classifier[2] to obtain such mapping rules. Another approach is to use Linear Multivariate Regression and Dynamic Frequency Warping methods[1]. It is also important to extract unique characteristics from the limited training speech data.

Our method differs from the previous ones in two points. First, an orthogonal vector space conversion technique is proposed to transform the LPC cepstrum coefficients. Assuming that one person's LPC cepstrum coefficients can be represented by a signal vector in his(or her) own orthogonal vector space with finite dimension, the transformation of LPC cepstrum coefficient is efficiently implemented by substituting the vector space of the source with that of the target.

Second, we use minimum mean-squre error coordinate transformation (MSECT)[5] to move all the points in the source vector space to desired points in the target vector space. These transformation parameters are obtained from the time-aligned source and target LPC cepstrum coefficeints.

In addition to the transformation of the LPC cepstrum coefficients, pitch contour modification is applied to transform prosodic information. This transformation makes it possible to mimic the suprasegmental structure of the target speech and obtain a more natural sound whose properties are more similar to the target speech. The pitch contour modification process consists of two parts, global pitch modification and local pitch contour transformation.

In section 2, we will describe overall structure of the proposed voice personality transformation method. In section 3, the proposed transformation method for LPC cepstrum coefficients and the prosodic transformation rule will be described. Section 4 will present experimental results. Finally, conclusions and future work will be presented in section 5.

2. OVERALL STRUCTURE

The block diagram of the proposed voice transformation algorithm is shown in Fig. 1, which consists of three parts: "analyzer", "transformer", and "synthesizer". Each frame of the source speech is parameterized by LPC cepstrum coefficients, an excitation spectrum and a pitch period by the analyzer. In this paper, one analysis frame consists of 256 digitized speech samples and shifts 64 samples every frame (at a sampling frequency of 8KHz).

In the transformer, vocal tract information represented by the LPC cepstrum coefficients and prosodic

information represented by the pitch contour are transformed by the predefined rules. These rules are constructed in the "learning" or "training" stage.

Finally, the transformed speech is synthesized from the transformed LPC cepstrum coefficients and linear scaled excitation spectrum using the synchronized overlap and add(SOLA)[6] method.

3. TRANSFORMATION

In this paper, any LPC cepstrum coefficients of one speaker are modelled as a signal vector which is represented by the linear combination of orthogonal basis vectors. These basis vectors are refered to as the principle components of a speaker's LPC cepstrum coefficients. Accordingly, the source speaker's LPC cepstrum coefficeints are represented by the linear combination of the source speaker's principle orthogonal vector set. The same strategy is also used for the target speech.

From this model, the LPC cepstrum coefficients are efficiently transformed by replacing the source speaker's vector space with the target speaker's one. Fig. 2 is an illustration of this scenario. To implement this process, two major steps are required. One is to obtain the principle orthogonal basis vectors from the source/target training speech, and the other is to build up the source-to-target mapping rule.

3.1 Extracting principle orthogonal vectors.

Let C_i^s denote a column matrix of the i-th LPC cepstrum vector for the source speech signal. Each component of this column matrix is an LPC cepstrum coefficient.

$$C_j^s = [\ c_j^s(1)\ c_j^s(2)\ \dots\ c_j^s(P)\ ]^T \tag{1}$$

T denotes matrix transpose, and P is the total number of LPC cepstrum coefficients. Assuming that the correlation matrix is positive definite, its eigendecomposition is given by

$$R^s = \frac{1}{N^s}\sum_{i=1}^{N^s}(C_i^s\ C_i^{sT}) = Q^s \Lambda^s\ Q^{sT} \tag{2}$$

where N_s is the total number of training LPC cepstrum vectors. Q is eigenvector matrix, and Λ is a diagonal matrix whose diagonal components are the eigenvalues of the autocorrelation matrix. Then, the principle orthogonal vector set $V^{(s)}$ is composed of eigenvectors whose eigenvalue is larger than a given threshold value λ_{th}.

$$V^{(s)} = \{e_1^{(s)}, e_2^{(s)}, \dots, e_N^{(s)}\} \tag{3}$$

Note that any eigenvector $e_i^{(s)}$ in the set $V^{(s)}$ has its eigenvalue λ_i satisfying $\lambda_i \geq \lambda_{th}$. The principle orthogonal vector set for the target speech is obtained from the above procedure (1)~(3) and the only difference is that the training LPC cepstrum vectors are those of the target speaker. Using this orthogonal vector set, any source/target LPC cepstrum vector is represented by the KL(Karhunen-Loeve) expansion form.

$$C_i^s \cong \sum_{n=1}^{N} s_n^i e_n^{(s)},\quad C_i^t \cong \sum_{m=1}^{M} t_m^i e_m^{(t)} \tag{4}$$

where N, M is the total number of principle orthogonal vectors for the source/target, and s_k^i, t_k^i are the transfomation coefficients of the k-th principle orthogonal vector. LPC cepstrum vectors represented by the above equation would cause a small "reconstruction error", because the principle orthogonal vector set does not contain all the eigenvectors. However, the simulation results show that omitting eigenvectors with sufficiently small eigenvalues(<0.001) yield only an unnoticeable amount of noise.

3.2 LPC cepstrum transformation rule

The goal of the transformation is to minimize the error between the transformed source LPC cepstrum vectors and the corresponding target LPC cepstrum vectors. In this paper, an optimal transformation rule is considered as moving all the vectors in the source vector space to desired points in the target vector space. This rule is obtained from the training stage.

Training words uttered by both the source and target speakers are first acquired. This corpus is then LPC-analyzed with a fixed frame rate. Each word spoken by the source speaker is then time-aligned with corresponding word pronounced by the target speaker, using Dynamic Time Warping(DTW).

Thereafter, the time-aligned source/target LPC cepstrum vectors are projected onto the source /target principle orthogonal vectors. The transformed coefficients of the source/target cepstrum are given by

$$s_n^j = \sum_{i=1}^{P} e_n^{(s)} c_j^s(i)\ ,\quad \text{for } 1 \leq n \leq N \tag{5}$$

$$t_m^j = \sum_{i=1}^{P} e_m^{(t)} c_j^t(i)\ ,\quad \text{for } 1 \leq m \leq M \tag{6}$$

Then, we use the minimum mean-square error coordinate transformation (MSECT)[5] which is realized by simple matrix operations. The formulation of this idea is:

$$\hat{t}_m^{\,j} = \sum_{n=1}^{N} h_{mn} s_n^j + o_m \quad \text{, for} \quad 1 \le m \le M \tag{7}$$

so that, $\hat{t}_m^{\,j}$ is an linear transformation of the s_n^j. We will find optimal transform coefficients h_{mn}^*, and o_m^* which minimize the mean square error ξ between the set of transformed coefficients and the set of time-aligned target coefficients.

$$\xi = \frac{1}{N_s} \sum_{j=1}^{N_s} \sum_{m=1}^{M} (\hat{t}_m^{\,j} - t_m^{\,j})^2 \tag{8}$$

The linear equations for the optimum solutions h_{mn}^*, and o_m^* are given by

$$\frac{\partial \xi}{\partial h_{mn}} = 0 \ , \quad \frac{\partial \xi}{\partial o_m} = 0 \tag{9}$$

These linear equations are easily solved by the orthogonal property of KL-transformation coefficients [4]. The proposed transformation process is depicted in Fig. 3. To further minimize ξ, the values for h_{mn}^* and o_m^* are separated into classes. This is achieved by clustering the source LPC cepstrum vectors, and calculating optimum solutions for each cluster. At the transformation stage, a given LPC cepstrum vector is first classified, and the corresponding h_{mn}^* and o_m^* are used. We use the LBG algorithm to cluster the source LPC cepstrum vectors. According to simulation results, we find that the optimal number of clusters is 32, when both performance and computational load are considered.

The final transformed LPC cepstrum coefficients are obtained by the linear combination of principle orthogonal vectors for target speaker.

$$\hat{C}^t = \sum_{m=1}^{M} \hat{t}_m e_m^{(t)} \tag{10}$$

3.3. Prosody Modification

The prosody information characterizes one speaker's speaking style. This includes the speak -ing rate, pitch contour, energy contour and dynamic characteristics of formant frequencies. In this paper, limit prosody information to the pitch contour. Therefore, pitch contour modification is used to transform one person's prosodic charact-eristics.

The pitch information for one person is divided into two parts, the global and local characteristics. The average pitch value of the entire training speech data is considered as a global characteristic. Global pitch modification is achieved by a simple pitch scaling process. The transformed pitch period is given by

$$\hat{p}^{(t)} = \beta p^{(s)}, \quad \beta = \frac{p_{ave}^{(t)}}{p_{ave}^{(s)}} \tag{11}$$

where $p_{ave}^{(t)}$, $p_{ave}^{(s)}$ are the average pitch values for the taget and source speakers, respectively.

The local pitch contour is termed the "prosodic segment", and is obtained from the labeling of prosodic boundaries. We use a simple labeling method which investigates energy dips and abrupt pitch variations. Pitch contour is then normalized to have zero mean, and set to have equal duration by an interpolation and decimation process. Vector quantization is then applied to cluster these time-normalized prosodic segments.

The mapping rule is obtained simply from the following Bayesian rule.

$$\hat{P}^{(t)}(i) = \arg\max_k \ p(P^{(t)}(k) | P^{(s)}(i)) \tag{12}$$

where p(|) denotes conditional probability density function and $P^{(s)}(i), P^{(t)}(k)$ are the i-th prosodic segment of the source speaker and k-th prosodic segment of the target speaker, respectively. This means that given the pitch contour of the source, the transformed local pitch contour is the target pitch contour which has the most frequent occurrence.

4. EXPERIMENTAL RESULTS

To evaluate the proposed voice transformation algorithm, we perform two experiments for several speakers. Speech data is obtained from 4 male persons. The 4 speakers will be referred to by their initials KKS, KHG, SJT, and YJH, respectively, in this paper. Experimental conditions are summarized in table 1.

Table 1. Experimental conditions

A/D conversion	8KHz, 16bit
LPC order	12
LPC cepstrum order	20
No. of training sentences	10
No. of training words	124
codebook size for LPC cepstrum	32
codebook size for pitch contour	16

Two voice transformation experiments are presented: KHG-to-KKS and SJT-to-YJH. The results of the two experiments are listed in Table 2. The cepstral distances between the transformed cepstra and the time-aligned target cespstra are decreased by 44.3% and 58.1% compared to those that are not transformed, even in the non -clustering case. When the cluster size is 32, the decrease show ratios of 80.2% and 63.4%. These

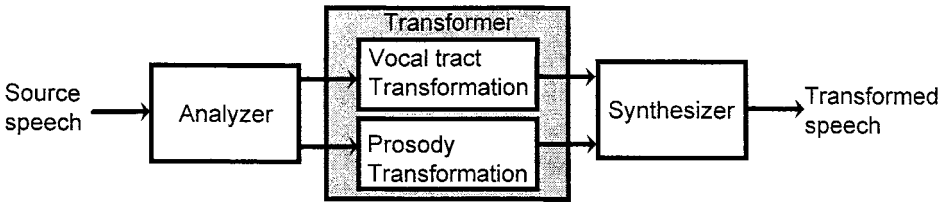


Fig. 1. Block diagram of the proposed voice transformation system

results confirm that the proposed LPC cepstrum transformation rule has a superior performance in an objective evaluation.

We also evaluated the effectiveness of the proposed method by conducting subjective listening

Table 2. Average cepstral distance

No. of cluster	KHG-to-KKS	SJT-to-YJH
no conversion	0.9941	0.8790
1	0.4402	0.5109
8	0.3391	0.3940
16	0.2822	0.3312
32	0.1965	0.3190

test which consist in presenting 3 utterences to 10 listeners. The first two utterances are the original source or target signals. The third one is the transformed speech. Listeners are asked to identify the speaker who might have pronounced the third utterance. The results of this experiment are shown in table 3. In KHG-to-KKS conversion, most of listeners can identfy transformed speech with target's one. However, the correct identification ratio is decreased to 69.1% in SJT-to-YJH conversion. It is caused by very different prosodic characteristics between the two speakers.

Table 3. Correct identification ratio

Experiment	correct identificaion ratio
KHG-to-KKS	82.7%
SJT-to-YJH	69.1%

5. CONCLUSION

We propose a new voice personality transformation technique through orthogonal vector space conversion and pitch contour modification. The proposed method suceed reasonably well in changing speaker personality, as proven by the average cepstral distance and subjective listening tests. In our method, prosody modification has been realized by changing only the pitch contour. To obtain perfect transformed speech, other prosodic features should be well modelled and transformed. This work remains as a future study.

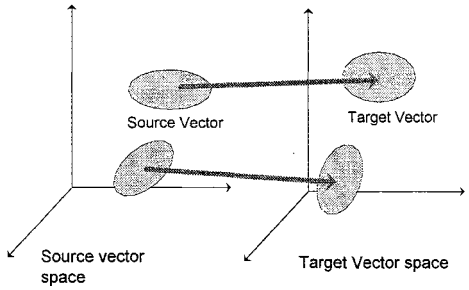


Fig. 2. Graphical explanation of vector space conversion.

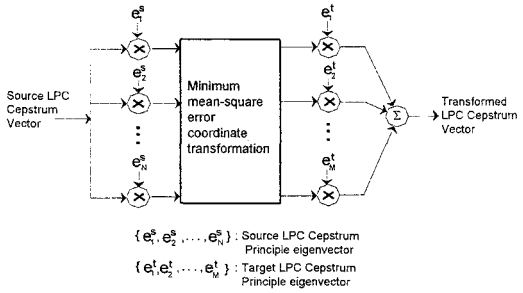


Fig. 3. Transformation of LPC cepstrum

References

[1] H.Valbret, E.Moulines, and J.P. Tubach (1992), "Voice transformation using PSOLA technique," Speech communication Vol. 11, pp. 175-187.

[2] Il Hyun Nam(1991), "Voice personality transformation," Ph.D. Thesis, Electrical Engineering Rensselaer Polytechnic Institute, Troy, New York.

[3] Y.Ephraim and H.L. Van Trees (1993), "A signal subspace approach for speech enhancement," Proc. ICASSP, Vol. 2, pp. 355-358.

[4] G. Strang (1980), "Linear Algebra and its applications," Academic Press Inc.

[5] S. A. Zahorian and A. J. Jagharghi (1992), "Minimum mean-square error transformations of categorical data to target positions," IEEE Trans. on Signal Processing, Vol. 40, No. 1, Jaunary, pp. 12-23.

[6] S. Roucos and A. M. Wilgus (1985), "High quality time-scale modification for speech," Proc. ICASSP Vol.1, pp. 493-469.