

VOICE CONVERSION BASED ON ACOUSTIC FEATURE TRANSFORMATION

Wei ZHANG, Liqin SHEN, Donald Tang
IBM China Research Laboratory, Beijing 100085, China
zhangzw, shenlq@cn.ibm.com

ABSTRACT

A voice conversion algorithm based on acoustic feature transformation is proposed. Concatenative speech synthesis is widely used nowadays. For such a system, voice conversion is much more challenging than a parameter based synthesizer, such as Klatt synthesizer. The main idea here is to construct transformations of the acoustic features between two speakers with Maximum Likelihood Linear Regression (MLLR) based on the acoustic feature domain concatenative speech synthesis. Some results of a series of experiments which are based on IBM trainable speech synthesis system, IBM cepstral reconstruction and LSP reconstruction algorithms are presented.

1. INTRODUCTION

For an advanced TTS system, speech conversion to different personalities based on different amount of available data is one of the most attractive aspects. In general, personality consists of two major factors: One is acoustic features and the other is prosodic features^[1]. According to previous studies, the acoustic features that contribute to speech personality are distributed among various parameters, such as formant frequencies, formant bandwidths, spectral tilt, and glottal waveforms^{[4][5]}. However there is no easy way to extract these parameters accurately then deploy them in synthesizing good quality speech. At certain level, some of the parameters are used in some parameter-based synthesizers. Consequently it is relatively easier to get some level of personality in such systems. On the other hand, many other approaches based on some computable acoustic parameters have been studied, for example, codebook mapping method by Abe et. al.^[1], probabilistic classification and HNM method by Yannis Stylianou et. al^[7].

Concatenative speech synthesis is widely used nowadays, because of its better voice quality and the availability of more powerful computing resources now. Normally it is based on a large amount of training corpus and the synthesized speech has the same personality of the original speaker. However it is not easy and practical to get so many corpus from everybody who wants to have a TTS system with his/her personality. It would be quite interesting to study the algorithms to implement voice conversion based on different amount of available data for a concatenative speech synthesis.

In this paper, we focus on the personality of acoustic features, and leave the personal prosodic features in the future study. Rather than using the normal time domain concatenative system, we developed an acoustic feature domain one, based on IBM cepstra reconstruction algorithm, also compared to the system based on the LSP parameters plus frequencies spectrum magnitude parameters of the residual. The general TTS system is trained by the source speaker, and the target speaker is the one whose personalized TTS is to be produced. The experiment is based on IBM trainable speech synthesis system^[2].

In the next section, the algorithms will be described. Following that in section 3, the experiment results are showed with some discussions. And the summary comes in Section 4.

2. ALGORITHM DESCRIPTION

2.1 IBM trainable speech synthesis system

With a large continuous speech corpus, IBM trainable speech synthesis system can automatically generate the alignment of speech segments corresponding to the synthesis units. Based on the decision tree algorithm, the context dependent acoustic tree can be constructed for each synthesis unit. Duration, pitch and energy values are predicted using the separate trainable models also. To determine the segment sequence to be

concatenated, a dynamic programming search is performed over all the waveform segments aligned to each leaf based on the defined cost functions. The selected segments are concatenated and modified to meet the required prosodic values using the FD-PSOLA algorithm. (Illustrated in Fig. 1)

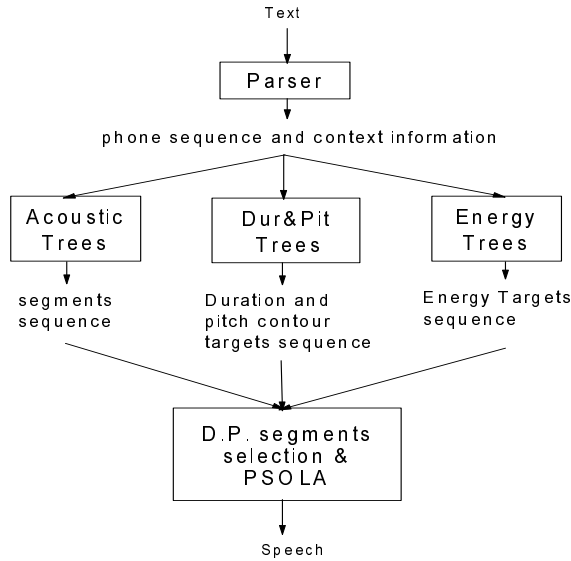


Figure 1 Diagram of Synthesis Procedure

2.2 Speech reconstruction algorithm base on Cepstra and LSP parameters plus frequencies spectrum magnitude parameters of the residual.

IBM Cepstral reconstruction algorithm^[3] is a novel one to reconstruct speech from the mel-frequency cepstral coefficients (MFCC) and pitch values.

LSP reconstruction algorithm is a traditional one to reconstruct speech from the LSP parameters, frequencies spectrum magnitude parameters of the residual and pitch values etc.

2.3 Voice conversion algorithm

For cepstral parameters and LSP parameters, the algorithm follows the 5 steps to construct the transformations for voice conversion.

Firstly, a defined amount of sentences are recorded by the target speaker. These sentences are part of the training scripts of our general TTS system.

Secondly, Using speaker-independent acoustic model to get the speech segment alignment results of these sentences.

Thirdly, for each segment in the target speakers' data, getting the corresponding one in the data of the speaker for the general TTS system and then traversing the latter in the acoustic parameters (CELP or LSP etc) related decision trees.

Then, if there is no enough data in some leaves, we can backtrack to the parent node until the data in the backtracked node is enough. And then we can get a series of Gaussian pairs about the acoustic parameters.

Finally, In the Gaussian pairs, mapping the data frame by frame, and then Maximum Likelihood Linear Regression (MLLR) is used to get these transformations for each pair. (Illustrated by figure 2)

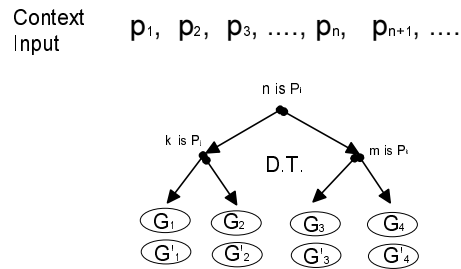


Figure 2. Transformations between the Gaussians of decision tree

After constructing the transformations, we can do voice conversion when synthesizing. When some segments are selected from the source speaker's speech database, we can traverse these segments in the decision trees through the context input and get the transforms. Each vector of acoustic parameters will be transformed to the approximation of the target speaker's one. Through the reconstruction algorithm, we can get the personalized speech.

In the algorithm, we propose to classify the phonetic related variances of the acoustic parameters by the Gaussians Clustering of decision tree, and approximate the variances of the acoustic parameters by MLLR.

3. EXPERIMENTS

We designed different transformation functions and experimented both cepstra features and LSP features.

3.1 Robustness of transformation functions and parameters

Let the X be the vector of the source speaker, Y be the target speaker's corresponding vector of, and Y' be the approximation of Y . Our object is to Let

$$Y' = T(X) \rightarrow Y.$$

we can evaluate

$$|\varepsilon| = |Y - Y'| = |Y - T(X)|$$

with different $T(*)$ functions.

We design four functions, $T_1(*)$, $T_2(*)$, $T_3(*)$ and $T_4(*)$

Let $T_1(*)$ be directly mapping function,

$$Y' = T_1(X) = X;$$

Let $T_2(*)$ be mean shift function,

$$Y' = T_2(X) = (X - X_m) + Y_m;$$

(X_m be the mean of X , and Y_m be the mean of Y)

Let $T_3(*)$ be the MLLR based matrix transform function between X and Y ,

$$Y' = T_3(X) = X * B;$$

(B is the MLLR transform matrix between X and Y)

Let $T_4(*)$ be the MLLR based matrix transform between the mean shift of X and Y (between $X - X_m$ and $Y - Y_m$)

$$Y' = T_4(X) = (X - X_m) * B_m - Y_m$$

(B_m is the MLLR transform matrix between $X - X_m$ and $Y - Y_m$)

The evaluation results of Cepstral parameters, LSP parameters and frequencies spectrum magnitude of residual are showed in Table 1, the values in the table are the average distances of each dimension:

	$ Y - T_1(X) $	$ Y - T_2(X) $	$ Y - T_3(X) $	$ Y - T_4(X) $
CELP	9.0477	6.0272	3.6623	3.4538
LSP	0.0466	0.0276	0.0187	0.0177
F.S.M	0.3567	0.3204	0.1744	0.1691

Table 1 : Average distances of parameters

From table 1, we can observe that the algorithm of Mean shift plus MLLR can obviously reduce the distance of $|Y - Y'|$. So $T_4(*)$ is selected to be the mapping function.

3.2 Spectrum analysis

We select a sentence in personalized training scripts and get 3 speech waveforms: the source speaker's original waveform, the target speaker's original waveform and the converted synthesized results (Illustrated by Fig3).

Although the spectrum of the converted synthesized sample is dimmer than the others, it seems that the spectrum of the last sample is more similar with the one of the second sample that the one of the first sample.

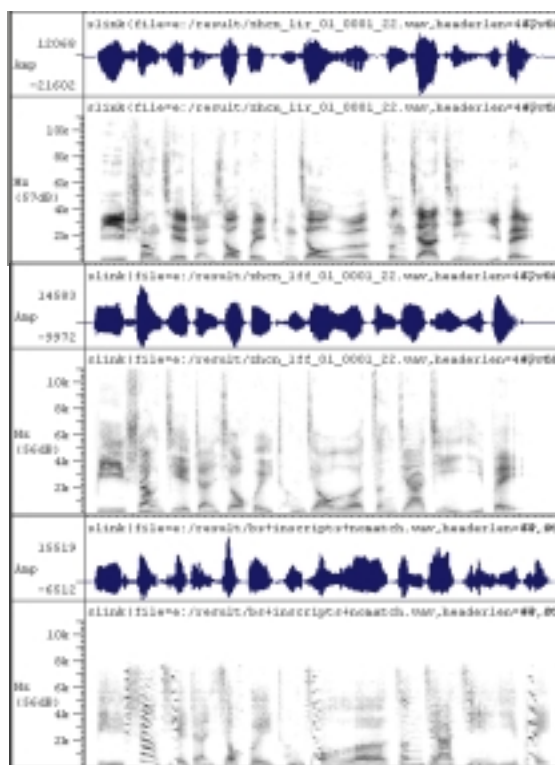


Figure 3 Spectrum of 3 samples

script "比上届展销会成交额增加近一倍"

1. The source speaker's original waveform
2. The target speaker's original waveform
3. The converted synthesized waveform

3.3 The listening experiment

Two sentences were synthesized by voice conversion, one is in the personalized training scripts, the other is out of them. 5 audiences evaluated these sentences. The source speaker is male and the target speaker is a female, so we design the listening experiment by asking two questions:

Question 1 : Is it female speech?

Answer : Yes, No.

Question 2 : Is it similar with the target speaker's speech?

Answer : Yes, No.

Question 3 : How about the quality of the speech?

Answer : Good enough (5), Good (4), Accepted (3), Bit accepted (2), Cannot accepted absolutely (1)

Sentences 1 is in the training scripts:

Liang2Tian1Zhang3Xiao1Hui4Cheng2Jiao1E2Da2Yi1Dian3Yi1Yi4Yuan2, Bi3Shang4Jie4Zhang3Xiao1Hui4Cheng2Jiao1E2Zeng1Jia1Jin4Yi2Bei4.

	Question 1	Question2	Question3
Cepstra	5:0	1:4	2.0
LSP+FSM	5:0	1:4	1.8

Table 2 : Listening result of sentence 1

Sentences 2 is not in the training scripts

Jiang1Ze2Ming2Shuo1, Ji4Hua4Sheng1Yu4He2Huan2Jin4Bao3Hu4Tong2Yan4Zhong4Yao4

	Question 1	Question2	Question3
Cepstra	5:0	1:4	2.0
LSP+FSM	5:0	1:4	1.8

Table 3 : Listening result of sentence 2

4. SUMMARIZATION

From the experiments and evaluation, especially from the answers of the first question, we can see that the voice conversion algorithm based on acoustic transformation is feasible. And the algorithm can be integrated with our IBM general TTS easily.

And the answers for the second question and the third question mean we have a lot of work to improve this algorithm. Firstly, we need improve the quality of the speech reconstruction results for our general feature

domain TTS system. Secondly, we need improve the clustering and regression algorithm. Thirdly, we should study prosodic feature transformation to make the converted speech more prosodically similar with the target speaker.

REFERENCE

- [1] M. Abe, S. Akamura, K. Shikano, and H. Kuwabara, Voice Conversion Through Vector Quantization, in Proc. IEEE ICASSP88
- [2] R. E. Donovan, and E. M. Eide, The IBM Trainable speech synthesis system, Proc. ICSLP98
- [3] D. Chazan, R. Hoory, M. Zibulski, Speech Reconstruction from the Mel-Frequency Cepstral Coefficients and Pitch Frequency. Proceedings of ICASSP, 2000
- [4] Kuwabara, H., Takagi, T., Quality Control of speech by Modifying Formant Frequencies and Bandwidth, 11th Inter. Congress of Phonetic Sciences, 1987
- [5] Childers, D. G. et. al., Voice Conversion: Factors Responsible for Quality, ICASSP85
- [6] C. J. Leggetter, P. C. Woodland, Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. Computer Speech and Language pp 171-185, Sept. 1995
- [7] Yannis Stylianou and Olivier Cappe, A System for Voice Conversion Based on Probabilistic Classification and a Harmonic Plus Noise Model.
- [8] Xiaochuan NIU, Liqin SHEN, Weibin ZHU and Qin SHI, Modelling and Decision Tree Based Prediction of Pitch Contour in IBM Mandarin Speech Synthesis System, ISCSLP 2000.
- [9] M. Gales, P. C. Woodland, Mean and Variance Adaptation within the MLLR Framework, Computer Speech and Language(1996)
- [10] Method for Generating Personalized Speech From Text. China Patent filed No. 01116305.4, Donald Tang, Liqin Shen, Wei Zhang, Qin Shi.