Improving the Performance of HMM-Based Voice Conversion using Context Clustering Decision Tree and Appropriate Regression Matrix Format

Long Qin[1], Yi-jian Wu[2], Zhen-Hua Ling[3], Ren-Hua Wang[4]

iFLYTEK Speech Laboratory
University of Science and Technology of China, Hefei, P.R.China
[1]qinlong@mail.ustc.edu.cn, [2]jasonwu@mail.ustc.edu.cn, [3]zhling@ustc.edu, [4]rhw@ustc.edu.cn

Abstract

To improve the performance of the HMM-based voice conversion system in which the LSP coefficient is introduced as the spectral representation, a model clustering technique to tie HMMs into classes for the model adaptation, considering the phonetic and linguistic contextual factors of HMMs, is adopted in this paper. Besides, due to the relationship between the LSP coefficients of adjacent orders, an appropriate format of the regression matrix is suggested according to the small amount of the adaptation training data. Subjective and objective tests prove that the source HMMs can be adapted more accurately using the proposed method, meanwhile the synthetic speech generated from the adapted model has better discrimination and speech quality.

Index Terms: model adaptation, regression matrix clustering, and regression matrix format

1. Introduction

With the development of the corpus-based speech synthesis technique, the intelligibility and naturalness of the synthetic speech has been improved a lot. However, it is still a difficult problem for the corpus-based TTS system to synthesize speech of various speakers and speaking styles with a limited database. So the voice conversion technique which can convert one speaker's voice to another speaker's voice provides a positive approach to achieve the goal of synthesizing speech of multi-speakers.

The HMM-based voice conversion system is built on the basis of the HMM-based speech synthesis. In the HMM-based speech synthesis system, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [1][2][3]. In addition, voice characteristics of the synthetic speech can be converted from one speaker to another by applying a model adaptation algorithm, such as the MLLR (maximum likelihood linear regression) algorithm [4][5], with a small amount of speech uttered by the target speaker.

We have realized a HMM-based speech synthesis system in which the LSP (line spectral pair) coefficients and the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectral contour) analysis-synthesis algorithm are introduced [6][7]. Then, by realizing the MLLR algorithm, we provide our synthesis system with the ability of synthesizing voice of various speakers. However, there still exist two main problems in the HMM-based voice conversion system. Firstly, the data-driven clustering method described in the MLLR algorithm ignores many contextual factors between HMMs, therefore some unrelated HMMs are forced into one class which will affect the accuracy of the model adaptation. Secondly, the system performance including the voice characteristics and voice quality of the synthetic speech decreases greatly when the adaptation training data is very limited. In order to solve these problems, a clustering method, considering the phonetic and linguistic connections between HMMs using the context decision tree, which has been applied similarly in both the HMM-based speech recognition and the HMM-based speech synthesis areas [8][9], is described in this paper. Moreover, an appropriate regression matrix format is suggested when very few training data is available, as the LSP coefficients of only several adjacent orders have strong correlations.

In the following part of this paper, an overview of our HMM-based voice conversion system is presented in section 2. Section 3 describes the details of the proposed context clustering decision tree and the appropriate regression matrix for the model adaptation. Section 4 presents the results of experiments including subjective and objective evaluations while section 5 provides a final conclusion.

2. Overview of HMM-based voice conversion

A framework of our HMM-based voice conversion system is shown in Figure 1. The system consists of three stages, the training stage, the adaptation stage, and the synthesis stage.

In the training stage, the LSP coefficients and the logarithm of fundamental frequency are extracted by the STRAIGHT analysis. Afterwards, their dynamic parameters including delta and delta-delta coefficients are calculated. The MSD (multi-space probability distribution) HMMs are introduced to model spectrum and pitch parameters because of the discontinuity of pitch observations. And state durations are modeled by the multi-dimensional Gaussian distributions.

In the adaptation stage, the spectrum, pitch and duration HMMs of the source speaker are all adapted to those of the target speaker. At first, the spectrum and pitch HMMs are adapted to the target speaker's HMMs by MLLR with the context decision tree clustering. Then, on the basis of the converted spectrum and pitch HMMs, the target speaker's utterances are segmented to get the duration adaptation data. So that the duration model adaptation can be achieved.

In the synthesis stage, according to the given text to be synthesized, a sentence HMM is constructed by concatenating the converted phoneme HMMs. From the sentence HMM, the LSP and pitch parameter sequences are obtained using the speech parameter generation algorithm, where phoneme durations are determined based on the state duration distributions. Finally, the generated parameter sequences of spectrum, converted from the LSP coefficients, and F0 are put into the STRAIGHT decoder to synthesize the target speaker's speech.

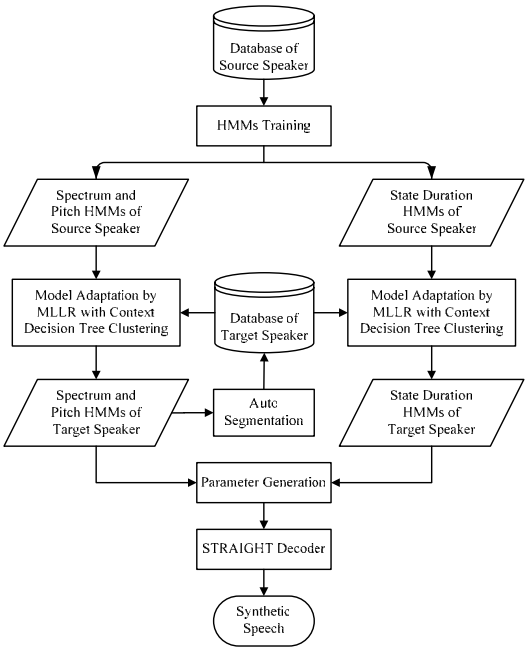

Figure 1: HMM-based voice conversion system

3. Context clustering decision tree and appropriate regression matrix format

3.1. Original MLLR algorithm

MLLR is a successful and widely used speaker adaptation algorithm in both speech recognition and speech synthesis areas. It estimates a series of linear transforms which can capture the differences between the source speaker's HMMs and the target speaker's HMMs under the ML (maximum likelihood) criterion. With the computed regression matrices, the parameters of the source speaker's HMMs can be transformed to those of the target speaker. But there isn't sufficient training data to calculate the regression matrix for each HMM. To overcome this problem, MLLR manages a data-driven clustering method to group HMMs into several classes in which all the models share the same regression matrix. Initially, all HMMs to be clustered are placed in the root node of the tree. This node is then split into two by maximizing the distribution distance between the two leaf nodes. The tree growth continues until each of the split nodes does not have sufficient training data to compute the regression matrix for HMMs in it. If a node does not have enough training data, HMMs will be modified by the regression matrix of its parent node.

Different from the speech recognition technique, there are many contextual factors between HMMs available in the speech synthesis system. But the data-driven clustering method ignores the context relations between HMMs which can not be reflected only by the distribution distances. As a result, some unrelated HMMs are forced into one class to share the same regression matrix. And that will cause two main problems. On one hand, HMMs can not be well adapted, as the regression matrix calculation for the tied HMMs is inexact. On the other hand, the speech quality of the synthetic speech is greatly affected by the errors generated due to the unreasonable clustering results. In order to solve these problems, we use the context decision tree clustering method instead of the regression class tree clustering method to tie HMMs during the MLLR model adaptation.

3.2. The Context decision tree clustering method

A context decision tree is a binary tree in which a question relating to the phonetic context to the immediate left or right is attached to each node as illustrated in Figure 2. The tree is built using a top-down sequential optimization procedure. Every time a node is split into two, the question which can partition HMMs in that node and give the maximum increase in the log likelihood of the tied HMMs is chosen. The node splitting procedure is repeated until the likelihood increase falls below a threshold which is calculated using the MDL (minimum description length) criterion. And to ensure that there is sufficient data associated with each terminal node to compute the regression matrix, an empirically derived minimum occupation count is applied. Many phonetic and linguistic contextual factors, such as the phoneme identity factors, the stress related factors and the locational factors, are taken into account to cluster HMMs. Consequently, the problems caused by the unrelated HMMs in a class are greatly reduced.

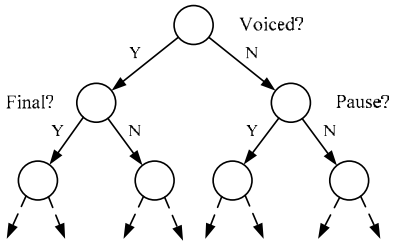

Figure 2: A context decision tree for clustering

3.3. Regression matrix format considering correlations between the LSP coefficients of adjacent orders

When the amount of the training data is very small, only a few regression matrices can be effectively calculated. Accordingly, the voice characteristics and voice quality of the synthetic speech degrade observably. However, if the regression matrix has fewer coefficients to estimate, more regression matrices can be generated and the adaptation will be more robust and accurate. Since the LSP coefficients of only a few adjacent orders have strong correlations, the elements of the spectral regression matrix which aren't near the diagonal are almost zero. So the block matrix, in which the diagonal elements are square matrices of any size and the off-diagonal elements are zero, and even the band matrix, which only has nonzero elements arranged uniformly near the diagonal, instead of the full matrix can be selected as the format of the regression matrix without great loss of adaptation precision. Moreover, as the block or the band regression matrix formats have fewer elements to calculate, the number of the clustered classes will increase, which means the more accurate HMMs adaptation and the better individuality of the synthetic speech. The best choice of the regression matrix format for the small amount of the training data will be discussed in the following experiments.

4. Experiments and evaluations

4.1. Experimental conditions

We collect 1100 phonetically balanced sentences of a male and a female speaker from a Chinese database. The two speaker's SD (speaker dependant) models are trained by 1000 sentences of them, respectively. And the rest 100 sentences are used for the model adaptation and evaluation. The speech is sampled at a rate of 16KHz. Spectrum and pitch is obtained by the STRAIGHT analysis. Then they are converted to the LSP coefficients and the logarithm F0 respectively, and their dynamic parameters are calculated. Finally, the feature vector of spectrum and pitch is composed of the 25-order LSP coefficients including the zeroth coefficient, the logarithm F0, as well as their delta and delta-delta coefficients. We use the 5-state left-to-right no-skip HMMs in which the spectral part of each state is modeled by the single diagonal Gaussian output distributions. The duration feature vector is a 5 dimensional vector, corresponding to the 5-state HMMs, and the state durations are modeled by the multi-dimensional Gaussian distributions.

4.2. Experiments on the context decision tree clustering

A female to male voice conversion is conducted using both the regression class tree clustering and the context decision tree clustering method. It is known that the value of the LSP coefficients increases as the order of LSP ascends. But we find that some LSP coefficients in the adapted HMMs are disordered, which means the LSP coefficients of the higher order is smaller than that of the lower order. Figure 3 shows the number of disorders in the adapted models when the different amount of the adaptation data is provided. The blue line corresponds to the experiment result with the regression class tree clustering method, while the red line gives the result with the context decision tree clustering method. It can be seen that the number of disorders decreases remarkably using the context decision tree clustering technique.

The *AEC* (adaptation effectiveness coefficient) which represents the adaptation effectiveness by comparing the synthetic speech respectively generated by the adapted model and target speaker's SD model can be calculated as below

$$AEC = \begin{cases} 0 & S \geq S_s \\ -\frac{1}{S_s - S_t} * S + \frac{S_s}{S_s - S_t} & S_t < S < S_s \\ 1 & S \leq S_t \end{cases} \quad (1)$$

where Ss is the average spectral distance between the target speech and the speech synthesized by the source speaker's SD model, St is the average spectral distance between the target speech and the speech synthesized by the target speaker's SD model, and S is the average spectral distance between the target speech and the synthetic speech using the proposed model adaptation method. The average spectral distance is calculated by averaging spectral distances between 20 corresponding sentences of the synthetic speech and the target speech. So if *AEC* is 0, it means the adaptation is useless. Contrarily, the source speaker's model is totally adapted to the target speaker's model, when *AEC* equals 1. From what is shown in Table 1, we can also find the improvement by tying HMMs with the context decision tree.

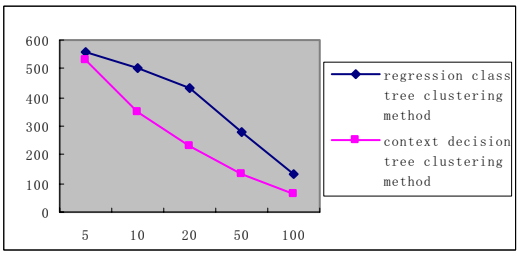

Figure 3: Comparison of the number of disorders in the adapted HMMs with the regression class tree clustering and the context decision tree clustering method

Table 1: Comparison of the adaptation effectiveness coefficient with the regression class tree clustering and the context decision tree clustering method

Amount of the training data	Regression class tree clustering	Context decision tree clustering
5	0.7026	0.7132
10	0.7597	0.7736
20	0.7803	0.8012
50	0.8215	0.8382
100	0.8468	0.8676

4.3. Experiments on regression matrix format selection

To find the optimal regression matrix format when the very limited adaptation training data is available, such as 5 or 10 sentences, comparative experiments using different regression matrix formats are performed. As Table 2 indicates, the most classes of the tied spectrum HMMs are collected when using the diagonal matrix, whereas the class number is the least using the full matrix.

Table 2: The number of the clustered classes using different regression matrix formats when only a few training data is available

Amount of the training data	Full matrix	3-block matrix	5-band matrix	Diagonal matrix
5	24	32	44	59
10	40	53	63	76

The number of disorders in the adapted HMMs also changes when the different regression matrix format is used. As presented in Figure 4, there are the most disorders in the adapted HMMs using the full regression matrix, while the number of disorders is the least when the diagonal regression matrix is executed. But the diagonal matrix is indeed not the optimal choice for the regression matrix, as it neglects all the relationship between the adjacent orders of the LSP coefficients which will also greatly affects the accuracy of the model adaptation. From the computed adaptation effectiveness coefficients in Table 3, we can find that the system performance is the best when choosing 5-band matrix as the regression matrix format. We also find out that when the training data is about 50 sentences to 100 sentences, the 3-block matrix is usually the appropriate option.

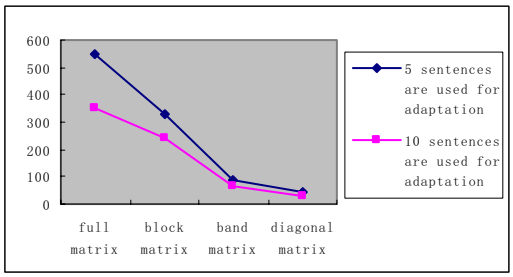

Figure 4: Comparison of the number of disorders in the adapted HMMs using different regression matrix formats when only a few training data is available

Table 3: The adaptation effectiveness coefficient using different regression matrix formats when only a few training data is available

Amount of the training data	Full matrix	3-block matrix	5-band matrix	Diagonal matrix
5	0.7132	0.7352	0.7529	0.7297
10	0.7736	0.8003	0.8250	0.7613

4.4. Experiments on speaker individuality and speech quality

The voice conversion experiment is applied between two speakers from both male to female and female to male. The first 1000 sentences are used for training the SD models, and 10 sentences stochastically selected from the last 100 sentences are used for the model adaptation which is realized by using the context decision tree clustering and the 5-band regression matrix. Ten listeners are asked to give the results of the listening tests. For comparing the proposed method and the data-driven method with the full regression matrix, we also list the performances of voice conversion using the conventional MLLR in brackets. And the MOS (Mean Opinion Score) test result of the synthetic speech using the target speaker dependant model is demonstrated as a reference.

Table 4: Subjective evaluation results

Conversion type	F to M	M to F	SD
Discrimination performance	4.44 (4.25)	4.56 (4.36)	---
MOS	2.80 (2.68)	2.89 (2.74)	3.15

Firstly, the synthetic speech is compared with the corresponding speech of the source speaker and the target speaker to get an evaluation based on the discrimination. We use 5 grades: 5 means very close to the target speaker while 1 is very close to the source speaker. The result is shown in Table 4. Secondly, in order to evaluate the quality of the synthetic speech, the opinion test is performed, where 5 means excellent and 1 means bad. The result is also shown in Table 4. Because the MOS of the speech synthesized by the target speaker's SD model is 3.15, there isn't much decrease in the procedure of the model adaptation.

5. Conclusion

The phonetic and linguistic contextual factors between HMMs are taken into account for the model adaptation in our HMM-based voice conversion system. At first a context decision tree is built under the ML criterion to group HMMs of the source speaker into several classes. And then both the mean and the variance regression matrices of each clustered class are estimated to convert the source speaker's HMMs to the target speaker's HMMs. Moreover, because of the strong correlations between the LSP coefficients of adjacent orders, an appropriate regression matrix format is suggested to improve the system performance due to the very limited adaptation training data. The results of subjective and objective tests indicate that the voice characteristics of the synthetic speech generated from the adapted model using the proposed method are closer to the target speaker than the conventional MLLR method. Meanwhile, the voice conversion system using the proposed method can synthesize speech with better speech quality.

6. Acknowledgement

This work was partially supported by the National Natural Science Foundation of China under grant number 60475015.

7. References

[1] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," *Proc. ICASSP-1996*, pp. 389-392, 1996

[2] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," *Proc. ICASSP-1999*, pp. 229-232, Mar. 1999.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration modeling for HMM-based speech synthesis," *Proc. ICSLP-1998*, vol.2, pp. 29-32, Nov. 1998.

[4] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol.9, no.2, pp. 171-185, 1995.

[5] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speaker adaptation for HMM-based speech synthesis system using MLLR," *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 273-276, Nov. 1998.

[6] Kawahara H., "Restructuring speech representations using a pitch-adaptive time frequency smoothing and a instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sound", *Speech Communication* 27, pp. 187-207, 1999

[7] Y.J. Wu and R.H. Wang, "HMM-based trainable speech synthesis for Chinese," *Journal of Chinese Information Processing*, accepted.

[8] S.J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," *Proc. ARPA Human Language Technology Workshop*, pp. 307-312, Mar. 1994.

[9] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," *Proc. ICASSP-2004*, vol.1, pp. 5-8, May 2004.