HIDDEN MARKOV MODEL BASED VOICE CONVERSION USING DYNAMIC CHARACTERISTICS OF SPEAKER

Eun-Kyoung Kim, Sangho Lee and Yung-Hwan Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology
373-1, Kusung-dong, Yusong-gu, Taejon, KOREA.
E-mail: ekkim@bulsai.kaist.ac.kr

ABSTRACT

This paper proposes a new voice conversion technique based on hidden Markov model (HMM) for modeling of speaker's dynamic characteristics. The basic idea of this technique is to use state transition probability as speaker's dynamic characteristics and have conversion rule at each state of HMM. A couple of methods is developed for creating state-dependent conversion rule. One uses source speaker's spectral dynamics and the other uses target speaker's. The experimental results showed that the proposed methods have better performance than conventional VQ-method in both objective and subjective tests. The comparison of our two methods showed that the method using target speaker's dynamics is superior in listening test and produces more natural sound.

1. INTRODUCTION

Voice conversion is a technique of transforming the characteristics of input speech in such a way that a human naturally perceives target's own characteristics in the transformed speech [1]. Among many potential applications of this technique, it is most important to create new voices for text-to-speech system using speech segment database.

Conventional voice conversion methods [2] make partitions of spectral feature space, and determines which partition's conversion rule is applied. They generally use static information of input frames for choosing conversion rule, and dynamic characteristics of speech are not considered. However, dynamic characteristics of speech can play an important role in determining partition of input speech since speech signal has time-varying property. This property is modeled by delta parameter or transition probability in speech recognition field. In voice conversion, it is also necessary that the dynamic characteristics of speech spectrum are modeled suitably for accurate and natural sounding spectral conversion [3].

We use HMM for modeling of dynamic characteristics. It is reasonable that each state has own conversion rule because the state of HMM is a set of acoustically similar feature parameters. And we can use state transitional probabilities as well as the information of frames for creating conversion rules at training phase. The outline of this paper is as follows. The structure of HMM for modeling of dynamic characteristics is presented in section 2. Two methods for generating conversion rules based on HMM are presented in section 3, and the performance of them is presented in section 4. Finally, conclusions are made in section 5

2. MODELING OF SPEAKER'S DYNAMIC CHARACTERISTICS

For modeling of speaker's dynamic characteristics, we use hidden Markov VQ model (HMVQM) [4] which is HMM having state-dependent codebook. In this model, VQ distortion measure using the state-dependent codebook is calculated instead of output probability at a state for the standard HMM. This is given by following equation.

$$b_i(X_t) = \exp(\max_k[-\mathrm{d}(X_t, C_k^i)]) \qquad (1)$$

Where, $\mathrm{d}(X_t, C_k^i)$ means distance between an input feature vector X_t and a codeword C_k^i of i-th state codebook. Therefore, output probability is the function of quantization distortion.

We modify HMVQM to perform spectral conversion. The model has two state-dependent codebooks which include spectral mapping relations between speakers. One represents partitioned acoustic space of source speaker and the other contains synthesis parameters for target voice. To simplify conversion process, we make two codebooks have one-to-one mapping relationship. The structure of HMM for conversion is a class of conversion rules with stochastic transitions as shown in Figure 1. We define this HMM as conv-HMM and two state-dependent codebooks as recognition-codebook and synthesis-codebook. In this paper, we will denote conv-HMM as λ=(A,RC,SC) which represent transition probability, recognition-codebook and synthesis-codebook. Similarly, HMVQM will be denoted by λ=(A,C) which are transition probability and state-dependent codebook. For the modeling of speaker's characteristics, we use ergodic type HMM.

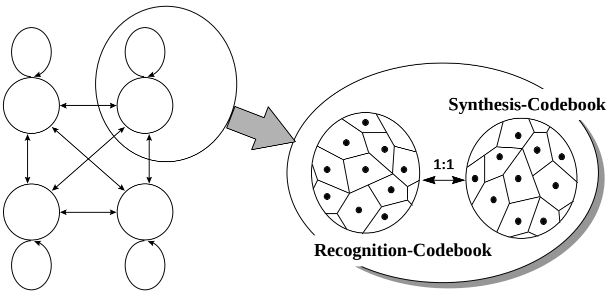

Figure 1. Structure of conv-HMM

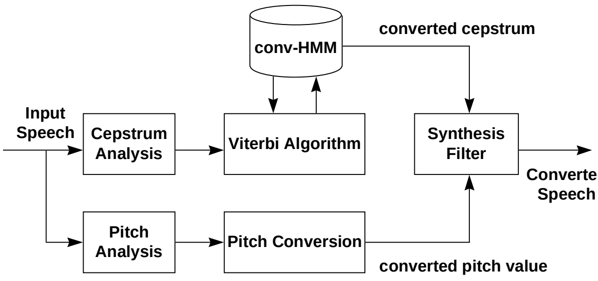

Figure 2. Block diagram of proposed system

The overall block diagram of proposed conversion procedure is shown in Figure 2. Input speech is parameterized by LPC cepstrum and recognized in the form of state sequence by Viterbi algorithm. After quantization process by recognition-codebook in each state, synthesis parameter is obtained by synthesis-codebook. Finally, synthesis filter produces converted speech with converted cepstrum parameters and converted pitch values as inputs.

3. CREATION OF CONV-HMM

Conv-HMM needs to include spectral mapping relation between two speakers. Therefore, procedure of creating conv-HMM is summarized as finding relationship between two speaker's parameters and generating conversion rule on each state. In this paper, two methods of creating conv-HMM are presented. The main difference of them is whose spectral dynamics are modeled. The first method uses source speaker's spectral dynamics as state transitional probabilities of conv-HMM. This method has the advantage of being able to select conversion rules considering source speaker's dynamic characteristics, but has drawbacks that converted speech have source speaker's spectral dynamics. Then, second method uses target speaker's spectral dynamics as transitional probabilities. In this method, converted speech has target speaker's spectral dynamics and can be expected of more natural sounding.

3.1. Using Source Speaker's Dynamics

Voice conversion function ϕ can be defined as finding the optimal sequence of target speaker's parameters given a sequence of source speaker's input parameters.

$$\phi(\mathbf{X}) \overset{\text{def}}{=} \arg\max_{\mathbf{Y}} P(\mathbf{Y}|\mathbf{X}) \qquad (2)$$

Where, $\mathbf{X} = \{X_1 X_2 \ldots X_T\}$ and $\mathbf{Y} = \{Y_1 Y_2 \ldots Y_T\}$ mean a sequence of source speaker's spectral parameters and target speaker's spectral parameters. Two parameters are time-aligned and T means length of them.

For a state sequence $S = s_1 s_2 \ldots s_T$ of input parameters on HMM, we rewrite conversion function ϕ as the following:

$$\begin{aligned} \phi(\mathbf{X}) &= \arg\max_{\mathbf{Y}} \sum_{S} P(\mathbf{Y}, S|\mathbf{X}) && (3) \\ &= \arg\max_{\mathbf{Y}} \sum_{S} P(\mathbf{Y}|S) \cdot P(S|\mathbf{X}) && (4) \\ &\cong \arg\max_{\mathbf{Y}} P(\mathbf{Y}|S^*) \cdot P(S^*|\mathbf{X}) && (5) \end{aligned}$$

Equation 4 is derived by assuming that **X** and **Y** are independent given S in equation 3, and equation 5 is derived by taking only the optimal state sequence $S^* = s_1^* s_2^* \ldots s_T^*$ into account, which is obtained by Viterbi algorithm. If we constrain that S^* is dependent on **X** then S^* can be assumed as a constant, and we rewrite the function ϕ` as the following:

$$\begin{aligned} \phi(\mathbf{X}) &= \arg\max_{\mathbf{Y}} P(\mathbf{Y}|S^*) && (6) \\ &\cong \arg\max_{\mathbf{Y}} P(\mathbf{Y}|RC_{k^*}^{S^*}) && (7) \end{aligned}$$

where, $RC_{k^*}^{S^*}$ is an optimal recognition-codeword sequence in S^*. To satisfy equation 7, we collect target speaker's training data on each optimal codeword of corresponding source speaker's training data. The synthesis-codebook are created using them. Therefore, the procedure of creating conv-HMM is as follows.

① Construct HMM $\lambda' = (A, C)$ which is trained by source speaker.

② Find an optimal state sequence S^* and quantize source speaker's training data to $RC_{k^*}^{S^*}$ on λ'.

③ Collect target speaker's training data on each codeword.

$$M_k^s = \{Y_t | Y_t \leftrightarrow X_t, quan(X_t) = RC_k^s\}$$ [1]

④ Synthesis-codeword consists of the mean of collected parameters.

[1] $Y_t \leftrightarrow X_t$ means that Y_t and X_t have correspondence by DTW, and *quan* function means quantization.

$$SC_k^s = \frac{1}{|M_k^s|}\sum_{x \in M_k^s} x$$

⑤ Define conv-HMM as $\lambda = (A, C, SC)$.

3.2. Using Target Speaker's Dynamics

Conversion function in equation 2 can be written as equation 8 by applying Bayes' rule.

$$\begin{aligned} \phi(\mathbf{X}) &= \arg\max_{\mathbf{Y}} P(\mathbf{X}|\mathbf{Y}) \cdot P(\mathbf{Y}) && (8) \\ &= \arg\max_{\mathbf{Y}} \prod_{t=1}^{T} P(X_t|Y_t) \sum_{S} \prod_{t=1}^{T} a_{s_{t-1}s_t} b_{s_t}(Y_t) && (9) \\ &\cong \arg\max_{\mathbf{Y}} \prod_{t=1}^{T} a_{s_{t-1}^* s_t^*} b_{s_t^*}(Y_t) P(X_t|Y_t) && (10) \end{aligned}$$

Equation 9 is derived by assuming that X_t is dependent on Y_t only and applying state transition probability and state output probability of HMM. It can be approximated by an optimal state sequence S^* as equation 10. By the way, the target speaker's parameters used for synthesis are confined to the codewords of synthesis-codebook. Therefore, the last term of equation 10 can be following equation.

$$\begin{aligned} P(X_t|Y_t) &\cong P(X_t|RC_k^{s_t}) P(RC_k^{s_t}|SC_k^{s_t}) && (11) \\ &\cong P(X_t|RC_k^{s_t}) && (12) \end{aligned}$$

$$\text{where, } k = \arg\max_i [-d(RC_i^{s_t}, X_t)]$$

Equation 12 is derived by assuming one-to-one mapping between RC and SC. Finally, conversion function ϕ can be rewritten as the following.

$$\phi(\mathbf{X}) \cong \arg\max_{\mathbf{Y}} \prod_{t=1}^{T} a_{s_{t-1}^* s_t^*} P(X_t|RC_k^{s_t}) \qquad (13)$$

To make recognition-codebook, we collect each source speaker's training data on each state. Also, process of reconstructing synthesis-codebook for one-to-one mapping should be considered. The procedure of creating conv-HMM is as follows.

① Construct HMM λ'=(A,C) which is trained by target speaker.
② Find optimal state sequence S^* of target speaker's training data on λ'.
③ Collect source speaker's training data on each state.

$$M^s = \{X_t | X_t \leftrightarrow Y_t, state(Y_t) = s\}$$

④ Create recognition-codebook RC^s by vector quantization of M^s.
⑤ Define synthesis-codebook SC as a linear combination of C weighted by the frequency of mappings between C and RC.
⑥ Define conv-HMM as $\lambda = (A, RC, SC)$.

4. EXPERIMENTS AND RESULTS

To evaluate the proposed voice conversion methods, we implemented two methods and a typical VQ-based method [6], and performed several experiments. We used histogram modification method for pitch conversion [7] and cepstrum synthesizer for converted speech. Experimental conditions are summarized in table 1.

Table 1. Experimental Conditions

A/D conversion	16kHz, 16bit
frame length	20ms
frame shift	5ms
feature parameter	20 order LPC cepstrum
codebook size (VQ)	512
configuration of HMM	32 state, 32 codeword

Speech data was obtained from 2 male and 2 female persons and each speaker will be referred to as m1, m2, f1, f2. Each speaker uttered 21 phonetically optimized words three times and 15 sentences two times. We carried out the experiments for eight of all possible pairs. The proposed HMM based conversion experiments are made using three configurations of HMM (16/32[2], 32/16, 32/32). As a result, 32/32 HMM was used for comparison test because it has similar quantization error with 512 codebook. We will denote two proposed methods as HMM_SRC and HMM_TAR and conventional method as VQ.

4.1. Evaluation By Cepstral Distance

We calculated average cepstral distance (CD) for objective evaluation of converted speech. Cepstral distance of one frame is defined as:

$$\mathrm{CD} = 10/\ln 10 \sqrt{2\sum_{i=1}^{p} (C_i^x - C_i^y)^2}\ (dB) \qquad (14)$$

where C_i^x and C_i^y denote i-th cepstrum parameter of converted and target speech and p means cepstrum order. Average cepstral distance can be defined as average CD value of time aligned converted speech and target speech. The results of test are shown in Figure 3. The proposed two methods are superior to typical VQ-based method for all experiment pair. Also, we can see that HMM_SRC method has the better result than HMM_TAR. It can be explained that conversion rule is trained for each codeword in HMM_SRC and then it has more accurate mapping function.

4.2. Evaluation By Listening Test

[2] number of state / codebook size

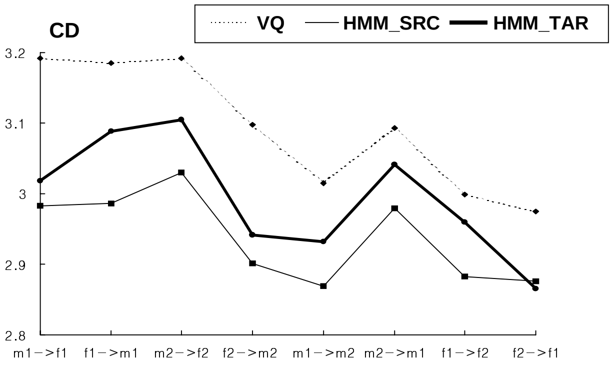

Figure 3. Cepstral distance of converted speech and target speech.

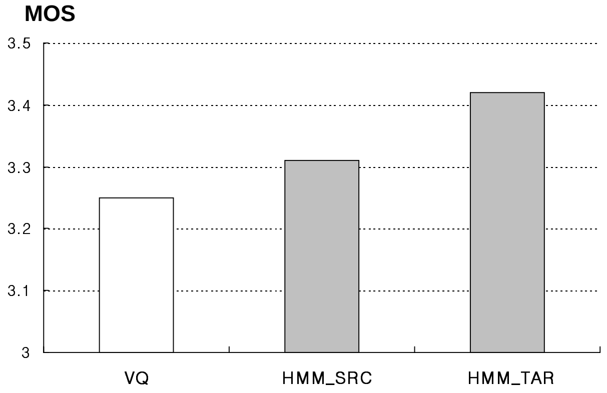

Figure 4. Result of MOS test.

To evaluate converted voice quality, subjective listening test was also carried out. Mean opinion score (MOS) test was adopted for calculating the similarity between converted speech and target speech. Listeners were asked to rate the similarity up to five degrees after listening to the converted speech and target speech. Twenty listeners participated in the experiment and they listened 8 sentences. The result is shown in figure 4, and it shows that the proposed methods are also superior in the sense of speech quality. We can find the interesting result that HMM_TAR gets higher score than HMM_SRC, which is the opposite result to objective test. It can be explained by figure 5 which shows that HMM_TAR have more accurate shape of spectral envelope. Therefore, it can be said that reflecting target speaker's dynamics have an effect on quality of converted speech.

5. CONCLUSION

In this paper, we have proposed HMM based voice conversion method using dynamic characteristics of speaker. Each state of ergodic HMM represents a subspace of speaker's acoustic space, so we modeled the dynamic characteristics of speaker by inter-state transitional probabilities and state-dependent conversion rule. Experimental results showed that the proposed systems are superior to conventional codebook mapping method based on vector quantization. And results also showed that the method of using target speaker's spectral dynamics is superior in listening test and produces more natural sound.

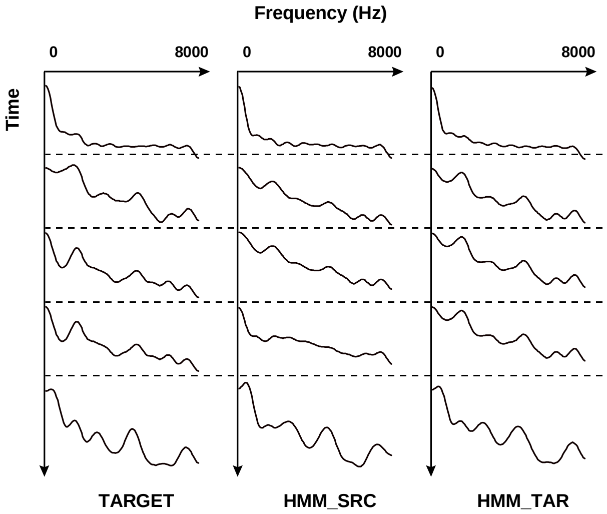

Figure 5. Spectral envelope of two methods

6. REFERENCES

[1] E. Moulines, Y. Sagisaka (eds.), *Voice Conversion : State of the Art and Perspectives*, Speech Communication, Vol.16, No. 2, pp. 125-126, 1995.

[2] G. Baudoin, Y. Stylianou, " On the transformation of the speech spectrum for voice conversion," *Proc. of ICSLP*, pp.1405-1408, 1996.

[3] Naoto Iwahashi, Yoshinori Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighing by radial basis function networks," Speech Communication, Vol.16, No.2, pp. 139-151.

[4] Seong Jin Yun, Yung Hwan Oh, "Performance Improvement of Speaker Recognition System for Small Training Data," *proc. of ICSLP*, pp1863-1866, 1994.

[5] X. D. Huang, Y. Ariki and M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[6] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice Conversion through Vector Quantization," *proc. of ICASSP*, pp.655-658, 1998.

[7] I. H. Nam, "Voice Personality Transformation," Ph.D. Thesis, Electrical Engineering, Rensselaer Polytechnic Institute, Troy, New York, 1991.