# HMM-Based Distributed Speech Synthesis Using Speaker-Adaptive Training

Kwang Myung Jeon and Hong Kook Kim

School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
{kmjeon, hongkook} @gist.ac.kr

Abstract. In this paper, a hidden Markov model (HMM)-based distributed text-to-speech (TTS) system is proposed to synthesize various speakers' voices in a client—server framework. The proposed system is based on speaker-adaptive training for constructing HMMs corresponding to each speaker, and distributes the operations of speech synthesis between a client and a server. That is, the speaker-adaptive training, which is a very complex operation, is assigned to the server, and less complex operations, such as text input and HMM-based speech synthesis, to the client. It is shown from performance evaluation that the proposed system operates in real time and provides good synthesized speech quality.

Keywords: HMM-based text-to-speech (TTS), distributed processing, client—server processing, speaker-adaptive training

## 1 Introduction

It is desirable that a text-to-speech (TTS) system be able to generate various kinds of voices. Such functionality can be realized by concatenative TTS, which is based on the concatenation of a speech segment from a set of large-scale speech databases for each type of voice. However, this approach is time- and cost-consuming for recording a large amount of speech data. As an alternative, a hidden Markov model (HMM)-based TTS with speaker-adaptive training can be employed for a limited data size for training a certain voice [1]. Despite the reduced data size, speaker-adaptive training is computationally expensive. Furthermore, it could be impractical to implement HMM-based TTS with such adaptive training on an embedded system with limited computational resources.

In this paper, we propose an HMM-based distributed TTS system in a client—server framework. In other words, speaker-adaptive training is performed on a server, while input/output operations, such as target voice recording, text inputting, and speech synthesizing, are performed at the client side. Thus, the proposed TTS system can offer synthesis of various speakers' voices by speaker-adaptive training in real time.

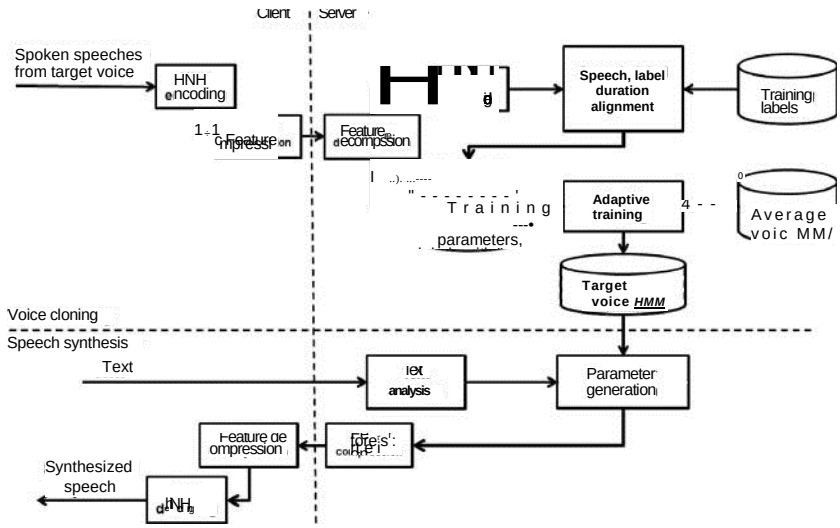


Fig. 1. Block diagram of the proposed HMM-based distributed TTS system

## 2 HMM-Based Distributed TTS System

Fig. 1 shows a block diagram of the proposed HMM-based distributed TTS system. As shown in the figure, the proposed system is divided into two stages: *voice cloning* and *speech synthesis.* In the voice cloning stage, the speech database (DB) of a target voice is recorded at the client side. While recording, the speaker whose voice is to be cloned is requested to speak predefined sentences that are 10 min long on average. It should be mentioned that the predefined sentences should be phonetically balanced to achieve efficient speaker-adaptive training. The obtained speech DB of the target voice is parameterized by harmonic and non-harmonic (HNH) modeling of speech [2]. In the HNH modeling, the number of parameters is 27, comprising 24 mel-frequency cepstral coefficients, FO, maximum voiced frequency, and gain ratio. Moreover, the parameters are compressed to have a bit rate of 8.3 kbit/s using a low-bit-rate feature vector compression scheme [3]. After the compressed parameters of the speech DB of the target voice are sent to the server, they are converted back to a speech signal through parameter decompression followed by HNH decoding. These reconstructed speech signals of the target voice are used to generate duration labels. Finally, the parameters of the target voice and duration labels are used to perform adaptive training applied to average voice HMMs [1]. In this paper, constrained structural maximum *a posteriori* linear regression [1] is used for the adaptation to prepare acoustic models of the target voice.

In the speech synthesis stage, the text input given at the client side is sent directly to the server. The text is then analyzed and converted to labels. The acoustic parameters corresponding to the labels are extracted from the acoustic models of the target voice prepared in the voice cloning stage using a parameter generation technique [4]. These acoustic parameters are compressed, transmitted back to the

client, and then decompressed through table look-up of the parameter codebooks [3]. Finally, the transmitted parameters are decoded into a synthetic speech signal.

## 3 Performance Evaluation

The performance of each stage of the proposed HMM-based distributed TTS system was evaluated. To this end, the client side of the proposed system was implemented using a laptop having a single-core processor with a clock speed of 1.6 GHz, while the server of the proposed system used a high-end workstation having a quad-core processor with a clock speed of 3.2 GHz. In addition, the client and the server were connected over a wireless network.

First, the processing time of the voice cloning stage was measured. Multiple trials of voice cloning processes were performed as follows. Eight participants uttered 100 predefined sentences each at the client side, which were used to train speaker-adaptive HMMs. In particular, the processing time of the speaker-adaptive training was measured as 9 min and 55 sec when the entire pronounced speech was 9 min and 38 sec. This implies that the real-time adaptation could be performed in the client—server framework.

Second, the performance of the speech synthesis stage was measured as a means of subjective quality evaluation. For the test, 20 utterances from each of the 8 different speaker-adaptive HMMs were prepared. Eight participants listened to the samples and then were asked to give a mean opinion score (MOS) ranging from 0 to 5 for the intelligibility of the synthesized speech and the similarity between the actual speeches and their synthesized versions. Consequently, the proposed system gave MOS values of 3.47 and 3.76 for intelligibility and similarity, respectively.

## 4 Conclusion

In this paper, an HMM-based distributed TTS system was proposed in a client—server framework in order to reduce the computational complexity of speech synthesis with speaker-adaptive training. In other words, speaker-adaptive training was performed on a server, while speech synthesis processing was performed on a client. It was shown from performance evaluation that real-time speaker-adaptive training could be carried out at the server and that the speech synthesized by the proposed system was comparable to the adapted speech with good intelligibility.

**Acknowledgments.** This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2012-010636).

## References


1. Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., Renals, S.: Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 17, 6 (2009) pp. 1208-1230.
2. Jeon, K. M.: Harmonic and non-harmonic modeling of speech for statistical parametric speech synthesis. Master Thesis, School of Information and Communications, Gwangju Institute of Science and Technology, (2012).
3. Ramaswamy, G. N., Gopalalcrishnan, P. S.: Compression of acoustic features for speech recognition in network environments. In: Proceedings of ICASSP, 2 (1998) pp. 977-980.
4. Tokuda, K., Kobayashi, T., Imai, S.: Speech parameter generation from HMM using dynamic features. In: Proceedings of ICASSP, 1 (1995) pp. 660-663.