# Average Voice Modeling Based on Unbiased Decision Trees

Fahimeh Bahmaninezhad, Soheil Khorram, and Hossein Sameti

Speech Processing Lab., Department of Computer Engineering,
Sharif University of Technology, Tehran, Iran

**Abstract.** Speaker adaptive speech synthesis based on Hidden Semi-Markov Model (HSMM) has been demonstrated to be dramatically effective in the presence of confined amount of speech data. However, we could intensify this effectiveness by training the average voice model appropriately. Hence, this study presents a new method for training the average voice model. This method guarantees that data from every speaker contributes to all the leaves of decision tree. We considered this fact that small training data and highly diverse contexts of training speakers are considered as disadvantages which degrade the quality of average voice model impressively, and further influence the adapted model and synthetic speech unfavorably. The proposed method takes such difficulties into account in order to train a tailored average voice model with high quality. Consequently, as the experiments indicate, the proposed method outweighs the conventional one not only in the quality of synthetic speech but also in similarity to the natural voice. Our experiments show that the proposed method increases the CMOS test score by 0.6 to the conventional one.

## 1 Introduction

The prevalent HSMM-based synthesis system, so-called speaker-dependent [1], employs one specific speaker's speech data for training the statistical model; later, the achieved model is used for synthesizing speech. This approach lacks acceptable quality when the training data is limited; therefore, for the sake of building a new voice naturally, collecting and preparing a tailored database is an essential task which leads to excessive waste of time and cost. On the contrary, speaker adaptive framework results in high quality output compared with the former one [2, 3] in small databases. The superiority of speaker adaptive system over speaker-dependent one is confirmed in [3, 4]. In fact, speaker adaptive system benefits from available multi-speaker corpus in order to compensate the restriction of target speaker's speech data. Hence, we proceed to make improvements in speaker adaptive framework.

Speaker adaptive synthesis system trains *average voice model* using multi-speaker speech database. Average voice model is a collection of speaker independent synthesis units. Thereafter, the adapted model is obtained by transforming the average voice model through adaptation algorithms and target speaker's speech data. Finally, synthetic speech is generated from the adapted model.

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 89–96, 2013.

In contrast to popular misconception, diverse contexts between different training speakers result in improper speaker adaptive synthesizer. In other words, if the set of sentences pertaining to a person in the train database are positively different from others regarding their context then the average voice model will have tendency toward a specific gender or speaker. Moreover, small amount of training data causes such drawbacks too [5]. Therefore, adapting the average voice model and synthesizing speech would affect the synthetic speech destructively.

In [3] the author presented a new technique for context clustering to solve the above-mentioned problem; however, this new method does not ensure the identical contribution of every speaker to every leaf of decision tree in context clustering. Thus, it is highly probable to have leaves with great domination of a specific gender or speaker.

In this study, we recommend another option to conflict the tendency of average voice model toward a little proportion of database. The proposed method guarantees equal participation of each speaker in every leaf of decision tree in context clustering, which in its own turn results in high quality synthesized speech in contrast to the conventional speaker adaptive synthesis system.

The rest of this paper is structured as follows. Section 2 describes the conventional speaker adaptive synthesis system. Our proposed system is introduced in Section 3. Experimental conditions and results are presented in Section 4, and concluding remarks and our plans for the future work are mentioned in the final section.

## 2 Speaker Adaptive HSMM-Based Speech Synthesis System

Many papers, such as [3, 6, 7], describe the conventional speaker adaptive HSMM-based speech synthesis system in detail. Here, we merely investigate the structure of this system generally. The schema depicted in Fig. 1 represents the main parts of the speaker adaptive system, namely training, adaptation, and synthesis.

Based on Fig. 1 speaker adaptive system and the well-known speaker-dependent approach [1] have two major disparities, 1) in speaker adaptive system, training data comprises speech from multiple speakers, contrary to one speaker used in the speaker-dependent system; 2) the adaptation phase must be carried out in speaker adaptive system between training and synthesis steps to construct the specific target speaker's model.

In the training phase of speaker adaptive framework, spectral and excitation parameters are first extracted for each frame; then, a speaker-independent context-dependent [8, 9] multi-space-distribution HSMM [10] is modeled employing a set of training data collected from different speakers. The trained model of this step is a statistical model called *average voice model*.

Afterwards in adaptation step, the average voice model is transformed and adapted to the target speaker employing adaptation data and adaptation algorithms [11–15]. Finally, the output speech is synthesized through the adapted model in synthesis phase.

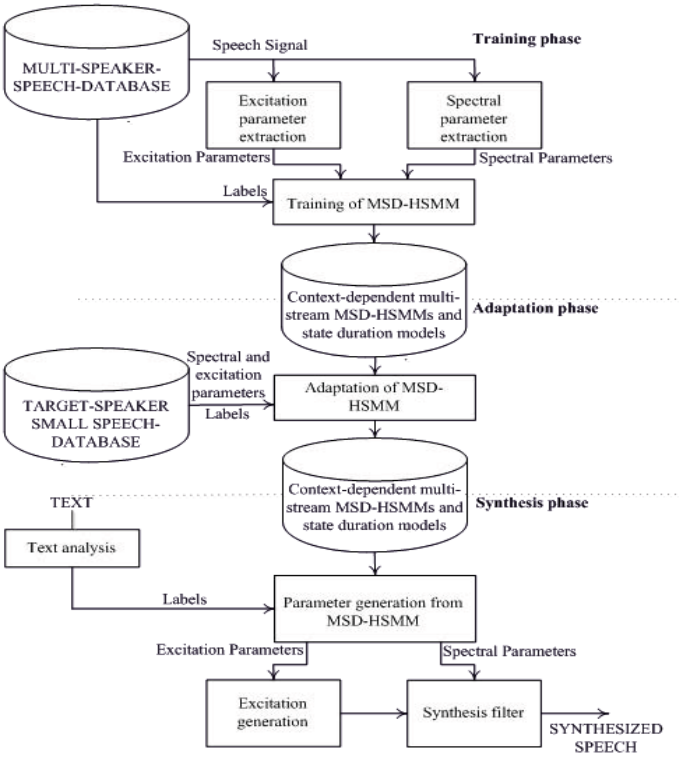



**Fig. 1.** Overview of speaker adaptive speech synthesis system [3]

All in all, the significant importance of speaker adaptive system is the necessity of very limited adaptation data. This effectiveness originates from the presence of rich and large training dataset. However, achieving a decent synthetic speech requires a high quality average voice model which is remarkably affected by the training data.

When the training corpus is small and embraces diversity among speakers' contexts we expect notable degradation in the quality of average voice model. Therefore, in such situations the widely-used decision-tree-based context clustering in training step causes an extreme reduction in the quality since average voice model would be biased toward just a limited section of database. In the next section we propose a new approach for training the average voice model which remarkably enhances the quality of average model.

## 3 Proposed Average Voice Model Training

In general, the average voice model affects the resulting adapted model and synthetic speech impressively; therefore, more favorable output will be synthesized if the average voice model is trained exactly and efficiently. Accordingly, our

contribution is solely performed to improve the effectiveness of the average voice model and is introduced in the current section.

Training database, embracing multi-speaker speech data, has a direct relationship with the quality of average voice model. In other words, large training database with relatively similar contexts among all speakers causes a decent average voice model; and other conditions lead to unfavorable average voice model. Therefore, we changed the conventional training phase to be compatible with every database. As a result, for any target speaker the synthetic speech would be more desirable in the proposed method.

More specifically, a training corpus with diverse contexts between various speakers leads to biased decision trees in the training step. In other words, each leaf of the tree may include data just from one speaker or one gender. This tendency which can govern a large number of leaves in the decision tree leads to degradation in resulting average voice model in conventional framework. Thus, the proposed method makes the decision tree contain data from all the training speakers.

Fig. 2 shows a general overview of the proposed training stage; the adaptation and synthesis phases are identical to the conventional structure. The main goal of this proposed model is enhancing the quality of output through the exact and better training of the average voice model without being biased toward any specific gender or speaker. The thorough explanation of the proposed training phase is given, hereinafter.

The proposed training procedure of average voice model could summarize as follows:

1) A speaker-dependent model is trained for each speaker in the training database [1].

2) A list of all contexts, i.e. the union of all speakers' contexts, is prepared.

3) For each speaker, the model of all contexts (full-context list), listed in the preceding step, are determined. We use the decision trees resulting from pre-trained speaker-dependent models to establish each context model.

4) The general model of every context in the full-context list is achieved by averaging among each speaker's model; in other words, expectation of means and variances among every speaker's model for a particular context leads to the mean and variance of general model for that specific context.

5) The achieved full-context general model is clustered by utilizing decision-tree-based context clustering. This clustered model is our proposed average voice model, and it is not biased toward a minor section of database anymore.

Eventually, we transformed this proposed average model to a specific model for the target speaker; this transformation employs adaptation algorithms and we have utilized Maximum Likelihood Linear Regression (MLLR) algorithm [12] and Maximum A Posteriori (MAP) estimation [14, 15] for this purpose. In the

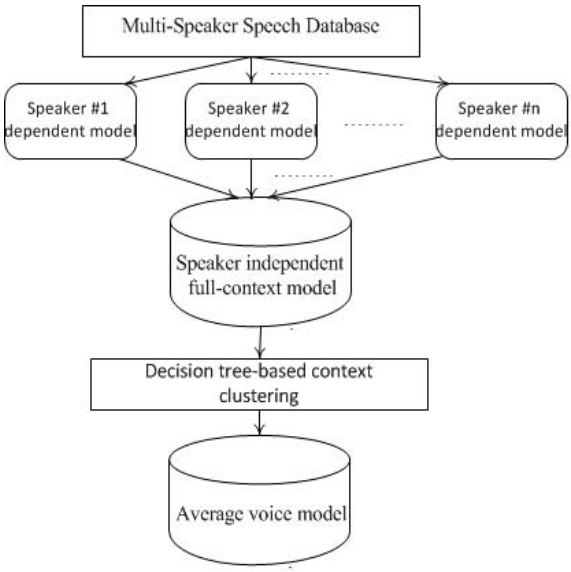



**Fig. 2.** Proposed method for training the average voice model

end, the favorable synthetic speech for the target speaker is resulted by applying the synthesis phase on the adapted model. These two steps are exactly alike those of conventional adaptation system [3].

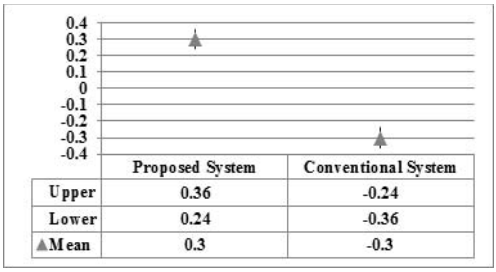
## 4 Experiments

### 4.1 Experimental Conditions

A Persian speech database, named FARsi Specch DATabase (FARSDAT) [16], is adopted to evaluate the proposed method. Several steps were carried out on FARSDAT to prepare it for the purpose of speaker adaptive speech synthesis. These steps in addition to FARSDAT characteristics were reported in our previous work [4].

The sampling rate of speech signals was 16 kHz, and they were windowed by a 25-ms Blackman window with a 5-ms shift. The feature vector embraced mel-cepstral coefficients (mcep), bandpass aperiodicity (bap), and fundamental frequency (log-F0) that were extracted by STRAIGHT [17]. In addition, 5-state left-to-right context-dependent HSMMs without skip paths were used. The synthesis units were modeled by considering segmental and supra-segmental contextual features; they are partly introduced in [4].

Four male and four female speakers' speech data, constituting the training corpus, were arbitrarily selected from FARSDAT to conduct the experiments, and a male speaker's speech data was selected for adaptation data. Training and adaptation data respectively consist of about 360 and 50 minutes. Moreover, it should be noted that there is not any overlap between training and adaptation




| | Proposed System | Conventional System |
|---|---|---|
| Upper | 0.36 | -0.24 |
| Lower | 0.24 | -0.36 |
| ▲Mean | 0.3 | -0.3 |

**Fig. 3.** Subjective evaluation of proposed system in contrast to conventional system based on CMOS test

data; and the adaptation was performed using MLLR adaptation [12] and MAP estimation [14, 15].

### 4.2 Experimental Results

This research takes the advantage of subjective test to evaluate the proposed method through two experiments. In both experiments, ten subjects were presented with seven synthesized speech. It is worthwhile to state that, seven test utterances were randomly chosen from 18 synthesized speech sentences which were contained in neither the training nor the adaptation datasets.

**Subjective Evaluation of Proposed Method in Contrast to Conventional One.** We determined the preference of proposed system over the conventional one regarding their voice characteristics by Comparison Category Rating (CCR) test [18] based on Comparison Mean Opinion Score (CMOS) scale in this experiment. Subjects were presented with a pair of synthetic speech from proposed and conventional system in random order and asked which one sounded better.

The results in Fig. 3 illustrate the priority of proposed speaker adaptive system. The figure shows the score with 95% confidence interval of the test. This enhancement in the quality of synthesized speech is the consequence of better training of average voice model; it also indicates the remarkable effect of average voice model on the synthetic speech.

**Subjective Evaluation of Synthesized Speech versus Natural Speech.** We examined the quality as well as the similarity of the synthetic speech to the target speaker's voice in current experiment; for this reason, we conducted Mean Opinion Score (MOS) test. Subjects scored the quality of synthetic speech generated from conventional and proposed system in the first test (Fig. 4). Additionally, they scored the similarity of speech synthesized from both systems to the original natural voice in the second test (Fig. 5).

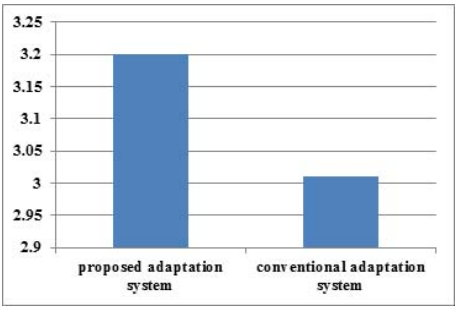



**Fig. 4.** Evaluation of quality

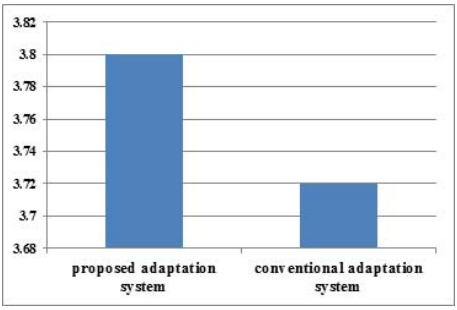



**Fig. 5.** Evaluation of similarity

As Fig. 4 shows the proposed system synthesizes speech with higher quality compared with the conventional one. Additionally, in Fig. 5 it is obvious that the speech synthesized from proposed system is more similar to the natural speech than that from conventional system.

## 5 Conclusion

In this research we proposed a new method for training the average voice model in speaker adaptive framework. The recommended system is more desirable against the conventional system when the training dataset is small or when speakers' contexts are very different. The subjective evaluations confirm the remarkable effect of the proposed average voice modeling on the quality of synthetic speech; therefore, the new output is more similar to the target speaker's voice and has higher quality in comparison with the conventional synthetic speech.

## References

1. Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Hidden semi-Markov model based speech synthesis. In: Proc. ICSLP, vol. 2, pp. 1397–1400 (October 2004)
2. Yamagishi, J., Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: A training method of average voice model for HMM-based speech synthesis. IEICE Trans. Fundamentals E86-A(8), 1956–1963 (2003)
3. Yamagishi, J.: Average-voice-based speech synthesis. Ph.D. thesis, Tokyo Institute of Technology (2006)
4. Bahmaninezhad, F., Sameti, H., Khorram, S.: HMM-based persian speech synthesis using limited adaptation data. In: 11th International Conference on Signal Processing (ICSP 2012), October 21-25. IEEE (2012)
5. Yamagishi, J., Masuko, T., Tokuda, K., Kobayashi, T.: A training method for average voice model based on shared decision tree context clustering and speaker adaptive training. In: Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), vol. 1, IEEE (April 2003)

6. Yamagishi, J., Ogata, K., Nakano, Y., Isogai, J., Kobayashi, T.: HSMM-based model adaptation algorithms for average-voice-based speech synthesis. In: Proc. ICASSP, Toulouse, France, vol. I, pp. 77–80 (May 2006)
7. Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Oura, K., Tokuda, K., Karhila, R., Kurimo, M.: Thousands of voices for HMM-based speech synthesis. In: Proc. Interspeech, pp. 420–423 (2009)
8. Shinoda, K., Watanabe, T.: MDL-based context-dependent subword modeling for speech recognition. J. Acoust. Soc. Japan (E) 21, 79–86 (2000)
9. Young, S.J., Odell, J.J., Woodland, P.C.: Tree-based state tying for high accuracy acoustic modeling. In: Proc. ARPA Human Language Technology Workshop, pp. 307–312 (March 1994)
10. Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T.: Multi-Space Probability Distribution HMM. IEICE Transaction on Information and Systems E85-D(3), 455–464 (2002)
11. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Trans. Audio, Speech, Lang. Process. 17(1), 66–83 (2009)
12. Leggetter, C.J., Woodland, P.C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9(2), 171–185 (1995)
13. Tamura, M., Masuko, T., Tokuda, K., Kobayashi, T.: Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In: Proc. ICASSP, pp. 805–808 (May 2001)
14. Lee, C.H., Lin, C.H., Juang, B.H.: A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models. IEEE Trans. Acoust., Speech, Signal Processing 39(4), 806–814 (1992)
15. Tsurumi, Y., Nakagawa, S.: An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum a Posteriori Probability Estimation. In: Proc. ICSLP 1994, S09-1.1, pp. 431–434 (1994)
16. Bijankhan, M., Sheikhzadegan, J., Roohani, M.R., Samareh, Y., Lucas, C., Tebiani, M.: The Speech Database of Farsi Spoken Language. In: Proc. 5th Australian Int. Conf. Speech Science and Technology (SST 1994), pp. 826–831 (1994)
17. Kawahara, H., Masuda-Katsuse, I., de Cheveign, A.: Restructuring Speech Representations Using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. Speech Communication 27(3-4), 187–207 (1999)
18. Recommendation ITU-U p.800, Methods for subjective determination of transmission quality. In: International Telecommunication Union (August 1996)