HMM-Based Persian Speech Synthesis Using Limited Adaptation Data

Fahimeh Bahmaninezhad, Hossein Sameti, Soheil Khorram
Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
bahmaninezhad@ce.sharif.edu, sameti@sharif.edu, khorram@ce.sharif.edu

Abstract—**Speech synthesis systems provided for the Persian language so far need various large-scale speech corpora to synthesize several target speakers' voice. Accordingly, synthesizing speech with a small amount of data seems to be essential in Persian. Taking advantage of a speaker adaptation in the speech synthesis systems makes it possible to generate speech with remarkable quality when the data of the speaker are limited. Here we conducted this method for the first time in Persian. This paper describes speaker adaptation based on Hidden Markov Models (HMMs) in Persian speech synthesis system for FARsi Speech DATabase (FARSDAT). In this regard, we prepared the whole FARSDAT, then for synthesizing speech with arbitrary speaker characteristics, we trained the average voice units; afterward, the adapted model was obtained by transforming the average voice model. We demonstrate that a few speech data of a target speaker are sufficient to obtain high quality synthetic speech, and we set out synthetic speech which has been generated from adapted models by using only 88 utterances is very close to that from speaker dependent models trained using 355 utterances.**

Keywords-component; Persian speech synthesis, average voice model, speaker adaptation, automatic annotation of FARSDAT

I. INTRODUCTION

Speech synthesis is referred to a technique which converts symbolic linguistic representations into human speech. A large variety of methods have been proposed, however HMM-based approach has dominated other methods over the last ten years [1, 2].

The main problem of building a new voice is to collect and prepare a labeled speech data. It is desirable to synthesize high quality speech using a small amount of speech data. This goal is achieved by employing speaker adaptation framework [3]. In other words, the modeling strategy is chosen by considering the amount of available speech data which belongs to the target speaker. Broadly speaking, speaker-dependent framework is an appropriate choice for a large amount of speech data while, adapting the average voice model to the target speaker [3] becomes favorable when available speech data of the target speaker are limited [4].

In speaker-adaptive framework, a large variety of contextual information extracted from several speakers' data, are utilized to build the average voice model. It provides a priori information for the speaker adaptation, and a robust basis is obtained as a result [5]. Thereby, the stable synthesized speech can be achieved even if the amount of speech data available for the target speaker is small.

Speaker adaptation is an issue of interest in most speech processing applications. Speaker adaptation techniques have been used in speech recognition systems [6] for quite a long time. Then, by introducing the HMM-based speech synthesis systems, these techniques are adopted to be used for speech synthesis. Speaker-adaptive HMM-based speech synthesis was initially implemented for Japanese language [4], and later was applied to English [7-10]. In this work, it is incorporated in Persian speech synthesis.

This paper describes the speaker-adaptive HMM-based speech synthesis framework for FARSDAT database [11]. FARSDAT is an Automatic Speech Recognition (ASR) corpus comprising utterances from 100 Persian speakers of different ages, sex, educations, and dialects. In this study, we extracted the features and contextual information of all the 100 speakers automatically, thus we could synthesize 100 speaker-adaptive HMM-based speech synthesis systems. Speech synthesis systems resulted from this work are valuable sources which can be used for further research areas such as eigenvoice conversion [12].

For generating each speaker's synthesis system, at first we model the average voice using train data of several speakers, then the adapted model is obtained by transforming the average voice model, using adaptation data of the target speaker, and finally synthesized speech are obtained from this adapted model.

The rest of the paper is organized as follows. Section 2 describes the Speaker-adaptive HMM-based speech synthesis system. In Section 3, database preparation is explained. Experimental conditions and results are described in Section 4, and concluding remarks and our plans for future work are presented in the final section.

II. SPEAKER-ADAPTIVE HMM-BASED SPEECH SYNTHESIS SYSTEM

As shown in Figure 1, the overall speaker-adaptive HMM-based speech synthesis framework consists of three stages namely: training, adaptation, and synthesis. In the following subsections, more details about these stages are given.

978-1-4673-2197-6/12/$31.00 ©2012 IEEE

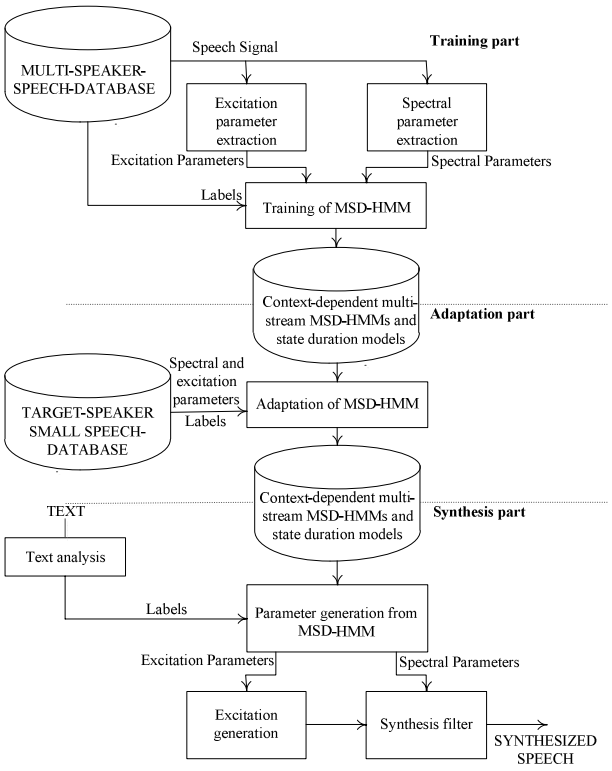


Figure 1. Overview of the average-voice-based speech synthesis system [9].

A. Training the average voice model

In this stage, context-dependent phoneme HMMs are trained by using speech data of several speakers, that is, training data. Trained context-dependent phoneme HMMs are obtained as follow.

We use spectral, excitation and state duration as parameters which are modeled for each context-dependent phoneme HMMs. Spectral parameters consist of mel-cepstral coefficients [13], their delta and delta-delta values, and for the excitation, the fundamental frequency consists of F0 logarithm in addition to its delta and delta-delta.

Firstly, spectral and F0 coefficients are extracted from the training speech dataset utterances and modeled by multi-stream HMMs. Output distributions for the spectral and F0 coefficients are modeled using a continuous probability distribution and multi-space probability distribution (MSD), respectively [2].

For each phoneme a context-independent HMM is trained as an initial model for corresponding context-dependent model. There are a large number of contextual combinations, and the speech dataset could not cover all of them; therefore, the decision-tree-based context clustering technique [14, 15] is applied separately to the spectral and F0 coefficients of the context-dependent phoneme HMMs. In the clustering techniques, a decision tree is constructed and similar models are tied based on predetermined criteria. Then re-estimation processes of the context-dependent phoneme HMMs are performed using the Baum-Welch (EM) algorithm. Finally, state durations are modeled by a multivariate Gaussian distribution [16], and the same state clustering technique is applied to the state duration models.

B. Adapting to the target speaker

In this phase, the average voice model, obtained from the training stage, is transformed and adapted to the target speaker using a small amount of speech data uttered by the target speaker. We can use a large variety of algorithms to achieve this goal, for instance, we can use the maximum likelihood linear regression (MLLR) algorithm [17] and MSD-MLLR algorithm [18] for spectrum and F0 adaptation, respectively.

C. Speech Synthesis

In the synthesis stage, first, an arbitrary given text is transformed into a sequence of context-dependent phoneme labels. In our work this is carried out by constructing a hierarchical tree for sentence utterance structure. Based on the label sequence, a sentence HMM is constructed by concatenating context-dependent phoneme HMMs. Spectral and F0 parameters of the sequence are generated from the resulting sentence context-dependent HMM, in which the phoneme duration are determined using state duration distributions. Finally, using an MLSA (Mel Log Spectral Approximation) filter [19, 20], speech is synthesized from the generated mel-cepstral and F0 parameter sequences.

III. DATABASE PREPARATION

For adapting Persian speakers, it is needed to consider special characteristics of this language, including utterance structure and the contextual factors. In this section details about FARSDAT and the pre-processing procedure are described.

The FARSDAT has been produced for speech and speaker recognition purposes in addition to linguistic research. Nevertheless, the above-mentioned method of synthesizing speech provides high quality voices by using these ASR corpora. Since FARSDAT includes a large number of speakers, this makes it possible to produce an enormous number of voices automatically. Consequently, this advantage enables us to improve the performance of significant applications such as eigenvoice conversion.

Each FARSDAT speaker has uttered roughly 300 unique utterances with the length of about 10 seconds in average. Manual phoneme-level labeling has been done for all utterances. FARSDAT sampling rate is 22.5 KHz, and its signal-to-noise ratio is 34 dB.

Since this database is not specifically designed for the text-to-speech (TTS) purpose, we had to go through certain pre-processing steps which, provides considered contextual information. In the rest of this section we briefly explain the FARSDAT preparations steps.

A. Extracting transcription

All utterances in FARSDAT were manually labeled both phonetically and phonemically. Hence, by considering this property we extracted the transcription of utterances using a Persian lexicon. This lexicon contains 60000 most commonly used words in the Persian. Afterward, for unspecified words, we used Peykare corpus [21]; this corpus is used because of the significant similarity that exists between Peykare and FARSDAT. Finally the remaining indeterminate words are specified manually.

B. Phoneme segmentation

The segmentation system employed here is based on the Hidden Markov Models Toolkit (HTK) [22]. Speech data of FARSDAT are used for modeling the phoneme HMMs. Therefore, MLLR (maximum likelihood linear regression) adaptation [17] scheme is used to adapt phoneme HMMs incorporating the speech data of each speaker. Finally, each speaker's adapted phoneme HMM is used for segmenting the data of that particular speaker. The segmentation is carried out by the Viterbi algorithm [23].

This method provides phoneme boundaries of FARSDAT, with a remarkable accuracy.

C. POS tagging

The most satisfactory approaches to specify Part Of Speech (POS) tags are HMM-based; therefore, we applied this parametric method. In Persian there are 25 different tags for words.

D. Specifing the stress pattern

Since the stress pattern in the Persian language is almost regulated, we simply used the specified POS tags and determine the stress pattern of all the utterances in the FARSDAT automatically.

IV. EXPERIMENTS

A. Experimental conditions

Speech signals were resampled at a rate of 16 KHz and windowed by a 25-ms Blackman window with a 5-ms shift. The feature vector consists of mel-cepstral coefficients, log fundamental frequency, and their delta and delta-delta parameters. The mel-cepstral coefficients were obtained from speech signal using a mel-cepstral analysis technique [20]. 5-state left-to-right context-dependent HMMs without skip paths were used.

Utterances from FARSDAT were used for training and adaptation. The average voice model HMMs were trained using arbitrarily-chosen four male and four female speakers' speech data, and the adapted model HMMs were achieved using one male speaker speech data as the target speaker.

The average voice model was trained using about 300 utterances for each training speaker. The average voice model was then adapted to the target speaker using the adaptation data whose utterances were not included in the training data. The adaptation was performed using MLLR adaptation [17] and MAP (maximum a posteriori) estimation [24, 25]. MLLR and MAP estimation were carried out sequentially for transforming the average voice model.

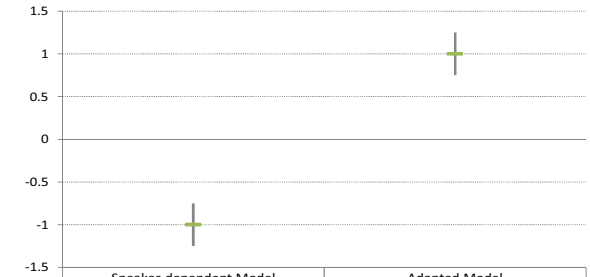
In the modeling of the synthesis units, the phonetic and linguistic contexts were taken into account. The Persian phonetic and linguistic contexts employed contain phonetic-level, syllable-level, word-level and utterance-level features. Some of the contextual factors that are considered are as follow:

- Phoneme
  - Preceding, current and succeeding phoneme.
  - Position of current phoneme in current syllable.
- Syllable
  - Stress of preceding, current and succeeding syllable.
  - Position of current syllable in current word.
  - Type of current syllable (syllables in the Persian language may be structured as CV, CVC, or CVCC).
- Word
  - Part Of Speech (POS) of preceding, current and succeeding word.
  - Position of current word in current utterance.
  - Current word contains "Ezafe" or not (Ezafe is a special feature in Persian which is the short vowel "e" and placed between two words such that it is not written but pronounced).
- Utterance
  - Number of syllables in current utterance.
  - Number of words in current utterance.
  - Type of current utterance.

B. Experimental Results

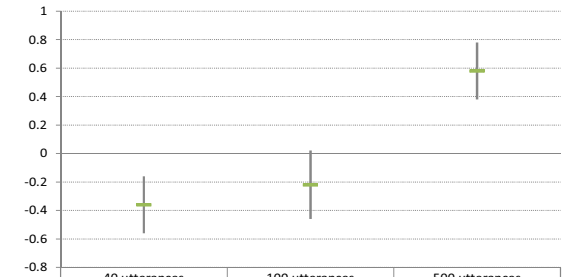
The above-mentioned speech synthesis method is evaluated in this subsection. Hence, two subjective tests are proposed. In the first test, we compared the quality of the synthesized speech, which is generated from the adapted models, with the synthesized speech resulted from the speaker-dependent model. On the other test, the comparison is between the synthesized speech of the adapted models when the number of adapting data are changing.

In both experiments, the quality of the adapted model is determined by a paired comparison test. Subjects were ten persons who were presented with pairs of synthesized speech from different models in random order and then were asked about their preference. For each subject, seven test utterances were chosen randomly, out of eighteen test utterances which were contained in neither the training nor the adaptation data utterance sets. We then conducted comparison category rating (CCR) test [26] to evaluate the effectiveness of synthesized speech from the adapted model. In CCR, the qualities of the



	Speaker-dependent Model	Adapted Model
Upper	-0.75	1.25
Lower	-1.25	0.75
Mean	-1	1

Figure 2. CCR test results for the speaker-dependent and adapted model.



	40 utterances	100 utterances	500 utterances
Upper	-0.16	0.02	0.78
Lower	-0.56	-0.4584	0.38
Mean	-0.36	-0.22	0.58

Figure 3. CCR test results for adapted model using different number of speech utterances from the target speaker.

pair of outputs from two different systems are scored by listeners using 7-point comparison mean opinion score (CMOS) scale [27]. The CMOS scale is a 7-point scale, that is, +3 for much better, +2 for better, +1 for slightly better, 0 for about the same, -1 for slightly worse, -2 for worse, and -3 for much worse.

1) Subjective evaluation of speaker adaptation method and speaker dependent method

We first evaluated the quality of the synthesized speech for the speech generated from the adapted model and the speaker-dependent model by a CCR test. In this experiment, 355 sentences of the target speaker are used for adapting the average voice model and training the speaker-dependent model.

Figure 2 shows the scores with 95% confidence interval of the test. From the result, we can see that adapting the average voice model to the target speaker becomes favorable when available speech of the target speaker is limited, comparing with the speaker-dependent model.

2) Subjective evaluation of speaker adaptation method with different number of adaptation data

We next evaluated the quality of the synthesized speech of the above-mentioned technique when the number of the target speaker' speech data is changing. At first, we used 355 utterances of the target speaker next we changed it to 187 utterances and at last to 88 utterances.

Figure 3 shows the result with 95% confidence interval of the test. The results show that by changing the number of adapting data, the quality of synthesized speech would remain unchanged, so by having a small number of data we can synthesize speech with satisfactory quality. In other words, by increasing the number of adapting data the quality would be slightly better but not significantly. In contrast, by reducing training data in speaker-dependent model the quality will reduce a lot.

V. CONCLUSION

In this paper, we have described the development and evaluation of the speaker-adaptive HMM-based Persian speech synthesis system. Persian linguistic information and contextual factors are considered in implementing this framework. From the results of the subjective tests, it is concluded that with a small amount of target speaker' data we can generate high quality synthesized speech with the adapted model. Our future work is enhancing the quality of synthesized speech by improving the average voice model.

ACKNOWLEDGMENT

This work was partially supported by Asr Gooyesh Company. Also, we want to thank Mr. Hossien Zeinali for his useful suggestions during the FARSDAT phoneme segmentation.

REFERENCES

[1] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis", in ICASSP, vol. 4, pp. 1229-1232, 2007.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. EUROSPEECH-99, pp. 2350-2374, September 1999.

[3] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis", IEICE Trans. Fundamentals, vol. E86-A, no. 8, pp. 1956-1963, Augest 2003.

[4] J. Yamagishi, "Average-Voice-Based Speech Synthesis", Ph.D. thesis, Tokyo Institute of Technology, 2006.

[5] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai, and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis", in Proc. ICASSP, Toulouse, France, vol. I, pp. 77-80, May 2006.

[6] M. Padmanabhan and D. Nahamoo, "Speaker clustering and transformation for speaker adaptation in speech recognition systems", IEEE Trans. Speech Audio Processing, vol. 6, no. 1, 1998.

[7] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis", IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 6, pp. 1208-1230, 2009.

[8] J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles", Proc. ICASSP'07, pp. 1233, 2007.

[9] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm", IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 1, pp. 66-83, January 2009.

[10] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila and M. Kurimo, "Thousands of voices for HMM-based speech synthesis", Proc. Interspeech, pp. 420-423, 2009.

[11] M. Bijankhan, J. Sheikhzadegan , M.R. Roohani, Y. Samareh, C. Lucas, and M. Tebiani, “The Speech Database of Farsi Spoken Language”, in Proc. 5th Australian Int. Conf. Speech Science and Technology (SST’94), pp. 826-831, 1994.

[12] T. Toda, Y. Ohtani, and K. Shikano, “One-to-many and many-to-one voice conversion based on eigenvoices”, Proc. ICASSP, vol. 4, pp. 1249, 2007.

[13] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech”, In: Proc. Of ICASSP, Minneapolis, Minnesota, USA, vol. 1, pp. 137-140, 1992.

[14] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition”, J. Acoust. Soc. Japan (E), 21:79-86, March 2000.

[15] S.J. Young, J.J. Odell, and P.C. Woodland, “Tree-based state tying for high accuracy acoustic modeling”, In Proc. ARPA Human Language Technology Workshop, pages 307-312, March 1994.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling for HMM-based speech synthesis”, In Proc. ICSLP-98, pages 29-32, December 1998.

[17] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, Computer Speech and Language, 9(2):171-185, 1995.

[18] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR”, In Proc. ICASSP 2001, pages 805-808, May 2001.

[19] S. Imai, “Cepstral analysis synthesis on the mel frequency scale”, In: Proc. Of ICASSP, Boston, Massachusetts, USA, pp. 93-96, Feb. 1983.

[20] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech”, In Proc. ICASSP-92, pages 137-140, March 1992.

[21] M. Bijankhan, J. Sheikhzadegan, M. Bahrani, and M. Ghayoomi, “Lessons from Creation of a Persian Written Corpus: Peykare”, Language Resources and Evaluation Journal, Springer Netherlands, vol. 45, pp. 143-164, 2010.

[22] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, O. Ollason, V. Valtchev, and P. Woodland, “The HTK Book”, Cambridge University, 2002.

[23] A.J. Viterbi, “Error bounds for convulutional codes and a asymptotically optimall decoding algorithm”, IEEE Transactions on Information Theory, 13:260-269, 1967.

[24] C.H. Lee, C.H. Lin, and B.H. Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models”, IEEE Trans. Acoust., Speech, Signal Processing, vol. 39, no. 4, pp. 806-814, 1992.

[25] Y. Tsurumi and Seiichi Nakagawa, “An Unsupervised Speaker Adaptation Method for Continuous Parameter HMM by Maximum a Posteriori Probability Estimation”, in Proc. ICSLP-94, S09-1.1, pp. 431-434, 1994.

[26] Recommendation ITU-U p.800, “Methods for subjective determination of transmission quality”, In: International Telecommunication Union, Augest 1996.

[27] V. Grancharov, and W. Kleijn, “Speech Quality Assessment”, Springer Handbook of speech processing chap. 5, 2007.