International Journal of Speech Technology (2019) 22:93–98
https://doi.org/10.1007/s10772-018-09578-2

Hidden-Markov-model based statistical parametric speech synthesis for Marathi with optimal number of hidden states

Suraj Pandurang Patil[1] · **Swapnil Laxman Lahudkar**[2]

Received: 30 January 2018 / Accepted: 27 November 2018 / Published online: 5 December 2018

Abstract
Hidden Markov Model and Deep Neural Networks based Statistical Parametric Speech Synthesis systems, gain a significant attention from researchers because of their flexibility in generating speech waveforms in diverse voice qualities as well as in styles. This paper describes HMM-based speech synthesis system (SPSS) for the Marathi language. In proposed synthesis method, speech parameter trajectories used for synthesis are generated from the trained hidden Markov models (HMM). We have recorded our database of 5300 phonetically balanced Marathi sentences to train the context-dependent HMM with five, seven and nine hidden states. The subjective quality measures (MOS and PWP) shows that the HMMs with seven hidden states are capable of giving an adequate quality of synthesized speech as compared to five state and with less time complexity than seven state HMMs. The contextual features used for experimentation are inclusive of a position of an observed phoneme in a respective syllable, word, and sentence.

Keywords Speech Synthesis · Hidden Markov Model · Context-dependent HMM · HMM Toolkit

1 Introduction

Over the last decade, speech synthesis research focus has moved from using traditional unit selection speech synthesis, where the small instances of speech are selected from large databases of natural speech, to a new methodology called Statistical Parametric Speech Synthesis (SPSS), where generative models are used to construct the speech waveforms (Black et al. 2012; Tokuda et al. 2002b).

Figure 1 shows the HMM-based Speech Synthesis system (HSPSS). The training part of HSPSS needs either the noise-free and well-labeled dataset or else the well-equipped pre-processing stage.

The noisy labels and overfitting are amongst the several factors which hamper intelligibility almost in all the speech enhancement and synthesis techniques, especially when it comes to capturing and recording real-time data. There are various approaches can be used for pre-processing to overcome this factors like active learning (Bouguelia et al. 2018) and nearest neighbour classifiers (Vajda and Santosh 2017) to accurately classify voice and unvoiced segments in case of noisy labels. Another direction of research that allows enhancing intelligibility of speech signal from noisy observations and multi-speech source recording is a direction of arrival estimation, source tracking and localization (Dey and Ashour 2018a, b, c). Training in HMM-based SPSS is similar to those used in Statistical Parametric Speech Recognition systems. The main difference is that both spectrum (e.g., Mel-cepstral coefficients (Fukada et al. 1992) and their delta and delta–delta features) and excitation parameters (e.g., log F0 and its delta and delta–delta features) are extracted from a speech waveforms and modeled by context-dependent (phonetic, linguistic, and prosodic contexts) HMMs are taken into account (Hunt and Black 1996). Multi-space probability distributions (Tokuda et al. 2002a) are used to model the log-fundamental frequency (log F0) sequence which includes unvoiced regions, for the state output stream for log F0. Each HMM has state duration densities to model the temporal structure of speech (Yoshimura et al. 1998). As a result, the system models spectrum, excitation, and durations in a unified framework (Black et al. 2007).

✉ Suraj Pandurang Patil
spatil.entc@gmail.com

[1] JSPMs Rajarshi Shahu College of Engineering, Pune, Maharashtra, India

[2] JSPM's Imperial College of Engineering and Research, Pune, Maharashtra, India

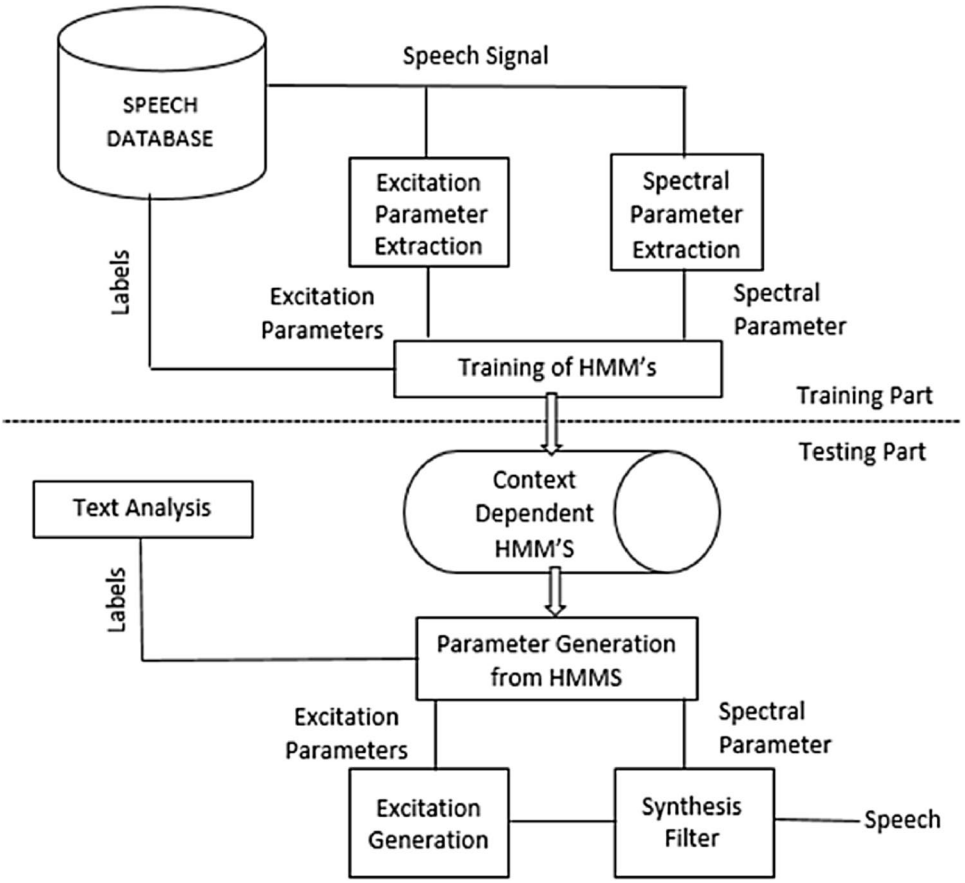

Fig. 1 HMM-based Speech Synthesis System

During testing/synthesis part, firstly, an arbitrarily given text corresponding an utterance to be synthesized is converted to a context-dependent label sequence and then the utterance HMM is constructed by concatenating the context dependent HMMs according to the label sequence. Secondly, state durations of the HMM are determined based on the state duration probability density functions (Imai 1983). Thirdly, the speech parameter generation algorithm generates the sequence of Mel-cepstral coefficients and log F0 values that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated Mel-cepstral coefficients and F0 values using the MLSA filter (Imai 1983) with a binary pulse or noise excitation.

This article gives the details of Marathi SPSS system based on HTS using Festival. Rest of the article is organized as follows: the chapter-2 contains the system overview, in chapter-3, we have explained the experimental setup and results. Chapter-4 Conclusion and Discussion gives the research outcomes from experimentations and results.

2 HMM-based SPSS system overview

2.1 Training part

During training of HMM-based SPSS systems, the HMMs are trained by applying input as linguistic feature vector and output vector of consists of normalized spectrum part and excitation part. In this work, the spectrum part consists of a 39-dimensional Mel-cepstral coefficient vector including the zeroth coefficients, their delta, and delta–delta coefficients. On the other hand, the excitation part consists of log fundamental frequency (log F0), its delta and delta–delta coefficients. HMMs have state duration densities to model the temporal structure of speech. As a result, HTS models not only the spectrum parameter but also F0 and duration in a unified framework of HMM.

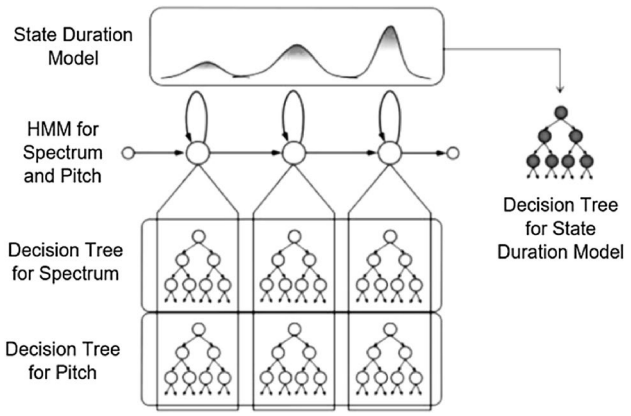

Fig. 2 Decision Trees for Context Clustering

2.1.1 Spectrum modeling

To control the synthesis filter by HMM, its system function should be defined by the output vector of HMM, i.e., Mel-cepstral coefficients. Thus we use a Mel-cepstral analysis technique, which enables speech to be re-synthesized directly from the Mel-cepstral coefficients using the MLSA (Mel Log Spectrum Approximation) filter (Fukada et al. 1992).

2.1.2 F0 modeling

The observation sequence of a fundamental frequency (F0) is composed of one-dimensional continuous values and discrete symbol which represents "unvoiced". Therefore, the conventional discrete or continuous HMMs cannot be applied to F0 pattern modeling. To model such observation sequences, we have proposed a new kind of HMM-based on a multi-space probability distribution (MSD-HMM) (Tokuda et al. 2002a). The MSD-HMM includes discrete HMM and continuous mixture HMM as special cases, and further can model the sequence of observation vectors with variable dimensionality including zero-dimensional observations, i.e., discrete symbols.

2.1.3 Duration modeling

State durations of each HMM are modeled by a multivariate Gaussian distribution. The dimensionality of state duration density of an HMM is equal to the number of states in the HMM, and the nth dimension of state duration densities is corresponding to the nth state of HMMs (Yoshimura et al. 1998).

2.1.4 Decision-tree based context clustering

The decision-tree based context clustering algorithm has been extended for MSD-HMMs (Yoshimura 2002). Since each of spectrum, F0 and duration has its own influential contextual factors, they are clustered independently as shown in Fig. 2 State durations of each HMM are modeled by an n-dimensional Gaussian, and context-dependent n-dimensional Gaussians are clustered by a decision tree. Note that the spectrum part and F0 part of the state output vector is modeled by multivariate Gaussian distributions and multi-space probability distributions, respectively (Tokuda et al. 2002b).

2.2 Synthesis part

In the synthesis part of HTS, first, an arbitrarily given text to be synthesized is converted to a context-based label sequence. Second, according to the label sequence, a sentence HMM is constructed by concatenating context-dependent HMMs (Tokuda et al. 2002b). State durations of the sentence HMM are determined so as to maximize the output probability of state durations, and then a sequence of Mel-cepstral coefficients and log (F0) values including voiced/unvoiced decisions is determined in such a way that its output probability for the HMM is maximized using the speech parameter generation algorithm. The main feature of the system is the use of dynamic coefficients: by the inclusion of dynamic coefficients in the feature vector, the speech parameter sequence generated in synthesis is constrained to be realistic, as defined by the statistical parameters of the HMMs. Finally, the speech waveform is synthesized directly from the generated Mel-cepstral coefficients and F0 values by using the MLSA filter (Tokuda 2006). Although a mixed excitation technique for HTS was already developed, the traditional excitation model was used in this work (Tokuda et al. 2002b).

3 Experimental setup and results

In this work, the following contextual factors are taken into account for the Marathi Language:

Phoneme:

- Preceding, current, succeeding.
- The position of a current phoneme in current syllable.

Syllable:

- Number of phonemes at {preceding, current, succeeding} syllable.

– The position of a current syllable in the current word.
– Vowel within current the syllable.

Word:

– Number of syllables in {preceding, current, succeeding} word.
– The position of a current word in a current phrase.
– Number of {preceding, succeeding} content words in the current phrase.
– The number of words {from previous, to the next} content word.

These factors are extracted from utterances using feature extraction functions of Festival speech synthesis system.

Acoustic features for experimentation

Here we have used a 39-dimensional Mel-cepstral coefficient vector including the zeroth coefficient, their delta, and delta–delta coefficients. On the other hand, the excitation part consists of log fundamental frequency (log F0), its delta and delta–delta coefficients.

The five, seven and nine state HMMs are modeled using these parameters respectively.

The HMMs described above have been trained on two different datasets recorded by native Marathi male and female speaker. Each dataset contains recordings of same 5300 phonetically balanced sentences totaling about 7.39 hours of speech per speaker. To obtain objective performance measures, 5100 sentences from each dataset were randomly selected as the training set for experimentation and remaining 200 sentences were used to form the test set.

All the three systems were built using an HTS, HMM speech synthesis toolkit version 2.3 on top of HTK 3.4.1.[1] The Mel-cepstral coefficients, the logarithm of F0 and aperiodic component features were modeled in separate streams during context-dependent HMM training. The duration of each HMM state is modeled by single Gaussian distribution. Once the spectral, F0 and aperiodic components had been estimated, a new set of state of alignments were computed and each state dependent Gaussian was estimated from state alignment statistics.

During the synthesis stage, global variance (GV) was used in the speech parameter generation algorithms to reduce the over-smoothing problem of HMM-based SPSS systems.

[1] http://htk.eng.cam.ac.uk/.

Table 1 RMSE for Acoustic Parameters generated using various number of Hidden States

Dataset	Number of hidden states	RMSE	
		Female	Male
Train	5	15.11	13.97
	7	10.89	10.03
	9	09.98	8.81
Test	5	15.55	13.07
	7	11.20	10.81
	9	10.88	10.32

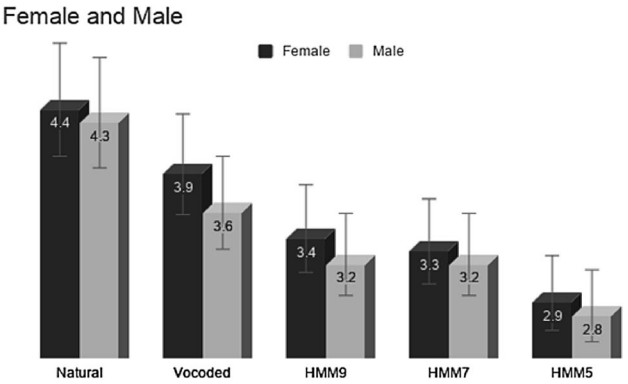

Fig. 3 Mean Opinion Score comparison chart

3.1 Objective tests

To quantitatively compare the three system, the root mean square error (RMSE) of Mel cepstral coefficients and log F0 observations were calculated for HMM5, HMM7, and HMM9 systems.

The Table 1 shows the RMSE obtained using HMM5, HMM7 and HMM9 systems.

It can be seen that HMM7 and HMM9 effectively reduce the RMSE in both the training and testing datasets. It also demonstrates that HMM9 outperforms HMM7 but at the cost of large computation and memory requirement. On the other hand, the HMM7 is reduced the RMSE nearer to HMM9 with comparative less computational overheads.

3.2 Subjective listening tests

Objectives measures are useful in comparing detailed system characteristics, the effective performance of speech synthesis systems can only be properly measured by conducting subjective listening tests. Here in this work, we have performed two listening tests:

Firstly, a Mean Opinion Score (MOS) test was conducted to compare the effectiveness of HMM5, HMM7, HMM9 systems for Marathi speech. Randomly, 20

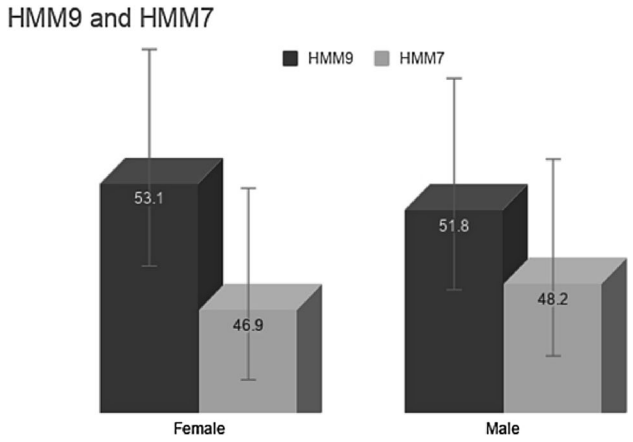

Fig. 4 Pairwise preference test for HMM with 9 and 7 hidden states

sentences were selected from the held-out tests sets and each listener was presented with ten sentences randomly selected from them of which 5 were male voices and 5 were female. The 10 native and 10 non-native listeners were asked to give the rating from 1 to 5 to each utterance. The rating definition was 1-bad to 5-excellent. In addition, natural and vocoded speech were also included in the test to determine the effects of vocoder artifacts on the assessment.

Figure 3 shows the results of MOS scores. It can be observed that the HMM9 and HMM7 systems outperform the HMM5 for both male and female voices.

Vocoded speech is the reference and best possible speech that can be synthesized using statistical models, as it is better than speech synthesized using HMM7 and HMM9 systems.

The second test was pairwise preference conducted to compare overall synthesis system performance. For this test also 20 test material sentences from newspapers were used. These sentences have the different pattern than that of dataset sentences and hence they provide a useful test with the generalization ability of the system. Two waveforms were synthesized for each sentence and each speakers using HMM7 and HMM9 systems respectively. Five sentences were randomly selected to make up a test set for each listener, leading to ten wave files pairs (5 male + 5 female). Each listener was asked to select the more natural utterance from each wave pair. The test results are shown in Fig. 4.

The results show that there is non-significant naturalness in pairwise preference test for HMM7 and HMM9.

Further, the comparison of synthesized speech from HMM-based SPSS system with seven hidden layers and current speech synthesis engines such as Sandesh Pathak and Dhvani shows the promising output.

4 Conclusion

We have applied an HMM-based speech synthesis system (HTS) to Marathi speech synthesis using Festival framework. In this work, we have evaluated and presented the results on HMM-based Marathi speech synthesis system with the various number of hidden layers. The results show that the system with seven hidden layers outperforms the system with five layers in terms objective metrics as well as perceptual preferences given by listeners. The synthesized speech of same system is also compared with already available speech synthesis engines like dhvani and Sandesh Pathak and results show that proposed system performed well as compared to existing speech engines.

Acknowledgements The Authors would like to thank Dr. K Samudravijaya for useful discussion on HMM-based Speech Synthesis Systems and his guidance for preparing and validating the prepared database. Authors also thankful to members of HTS working group including for their software development efforts.

References

Black, A.W., Bunnell, H. T., Ying, D., Muthukumar, P. K., Florian, M., Daniel, P., et al. (2012). Articulatory features for expressive speech synthesis. *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 4005–4008).

Black, A.W., Zen, H., Tokuda, K. (2007). Statistical parametric speech synthesis. *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Vol. 4, pp. 1229–1232).

Bouguelia, M. R., Nowaczyk, S., Santosh, K. C., Verikas, A. (2018 August). Agreeing to disagree: Active learning with noisy labels without crowdsourcing. *International Journal of Machine Learning & Cybernetics, 9*(8), 1307–1319.

Dey, N., & Ashour, A. S. (2018a). Applied Examples and Applications of Localization and Tracking Problem of Multiple Speech Sources. In N. Dey & A. S. Ashour (Eds.), *Direction of Arrival Estimation and Localization of Multi-Speech Sources, Springer Briefs in Electrical and Computer Engineering* (pp. 35–48). Cham: Springer.

Dey, N., & Ashour, A. S. (2018b). Sources localization and DOAE techniques of moving multiple sources. In N. Dey & A. S. Ashour (Eds.), *Direction of arrival estimation and localization of multi-speech sources, springer briefs in electrical and computer engineering* (pp. 23–34). Cham: Springer.

Dey, N., & Ashour, A. S. (2018c). Challenges and future perspectives in speech-sources direction of arrival estimation and localization. In N. Dey & A. S. Ashour (Eds.), *Direction of arrival estimation and localization of multi-speech sources, springer briefs in electrical and computer engineering* (pp. 49–52). Cham: Springer.

Fukada, T., & Tokuda, K., Kobayashi, T., Imai, S. (1992). An adaptive algorithm for mel-cepstral analysis of speech. *ICASSP* (pp. 137–140).

Hunt, A., & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *ICASSP* (pp. 373–376).

Imai, S. (1983). Cepstral analysis synthesis on the mel-frequency scale. *IEEE international conference on acoustics, speech and signal processing* (ICASSP) (pp. 93–96).

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T. (2002a). Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems, E85-D*(3), 455–464.

Tokuda, K., Zen, H., Black, A. W. (2002b). An HMM-based speech synthesis system applied to english. *IEEE workshop on speech synthesis*.

Tokuda, K. (2006). An HMM-based approach to flexible speech synthesis. In Q. Huo, B. Ma, E. S. Chng & H. Li (Eds.), *Chinese spoken language processing. Lecture notes in computer science* (Vol. 4274). Berlin: Springer.

Vajda, S., & Santosh, K. C. (2017). A fast k-nearest neighbor classifier using unsupervised clustering. In *Recent trends in image processing and pattern recognition. RTIP2R 2016. Communications in computer and information science* (Vol. 709). Singapore: Springer.

Yoshimura, T. (2002). Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-To-Speech systems, PhD dissertation, Nagoya Institute of Technology.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (1998). Duration modeling for HMM-based speech synthesis. *ICSLP* (pp. 29–32).