An HMM-based Cantonese Speech Synthesis System

Xin Wang
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University,
Shenzhen 518055, China
wangxin0624@gmail.com

Zhiyong Wu
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University,
Shenzhen 518055, China
zywu@sz.tsinghua.edu.cn

***Abstract*—This paper describes a Cantonese HMM-based speech synthesis system (HTS) using the general architecture of Crystal - a multilingual text-to-speech (TTS) framework developed in Tsinghua University. The generated synthesis engine of HTS has advantage of small footprint, the size of which is less than 7M bytes, and can be easily ported to embedded electronic devices such as smart-phones, set-top boxes, etc. Furthermore, the quality of the synthetic speech can be easily characterized by modifying the synthetic acoustic parameters of the proposed system. The result shows noticeable improvement in naturalness and smoother transition than the corpus-based unit-selection concatenative speech synthesis approach.**

Keywords—HMM model; Speech synthesis; Cantonese;

I. INTRODUCTION

In recent years, the corpus-based concatenative speech synthesis approach has been extensively used. It selects waveform units from a recorded large-scale speech corpus to generate natural synthetic speech. However, such approach lacks the ability in generating speech with both high voice quality and various voice characteristics such as expressive speech with different emotions and personalized speaking styles, etc. Furthermore, the usage of large corpus brings the difficulty in porting the speech synthesis system based on such approach to embedded electronic devices where the storage requirement is the core issue. To solve the problems, an HMM-based speech synthesis (HTS) approach has been presenting 0-[2], which trains a set of hidden Markov models (HMMs) from the training corpus and synthesizes speeches from the parameters of the trained HMMs. The acoustic features of the training corpus were parameterized by the HMM parameters. Hence the size of the footprint can be greatly reduced. Different voice characteristics can be easily achieved by training the HMMs from the data with different characteristics. In this paper, we apply the HMM-based speech synthesis system (HTS) to Cantonese using the general architecture of Crystal.

II. BUILDING CANTONESE SPEECH SYNTHESIS SYSTEM

There are two steps: training step and synthesis step to build an HMM-based Cantonese speech synthesis system. Figure 1 shows how this system works. The training stage consists of several core steps including data labeling and preprocessing, feature extraction, basic unit choosing for HMM models and selection of question set for constructing decision trees. During synthesis stage, HTS-engine [1]-[2] is used to generate the speech waveforms with the acoustic parameters generated from the trained HMM models.

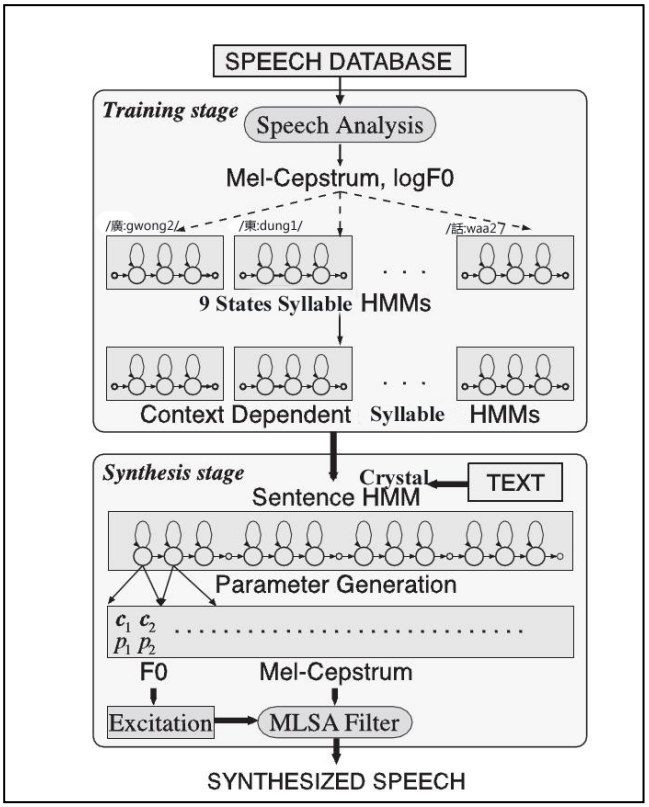


Figure 1. HMM-based speech synthesis system
(Adapted from [3] by using 9 states syllable HMMs)

A. Training Stage

In the training step, context dependent syllable HMM models are trained from a large corpus. The decision tree based context clustering approach is then used to train MFCC and logF0 HMM models. After that, the clustered HMM models are further re-estimated using the Baum-Welch (EM) algorithm, and the same method is applied to train the duration models except using multi-variate Gaussian distribution model [4]-[5].

This work is supported by the National Natural Science Foundation of China (60928005, 60805008, 60931160443), the Upgrading Plan Project of Shenzhen Key Laboratory and the Science and Technology R&D Funding of the Shenzhen Municipal.

1) Corpus and data preprocessing

The corpus contains 3000 sentences from the daily news, which is recorded by a native Cantonese female speaker and saved in Microsoft WAV format (16 kHz, 16 bit, mono channel). Each sentence contains exactly one prosodic phrase. 490 different tonal syllables were found in the corpus which covers almost all of the Cantonese initial-final pairs. The boundaries of the syllables for each sentence were labeled automatically with forced-alignment technique using a homegrown speech recognition engine, followed by manual check.

The used context label information include:

- The current syllable information (position, duration, stressed);
- The pronunciation information of the previous and next syllables (Jyutping, tone);
- The context information related to prosodic word and prosodic phrase (context information, prosody).

The above context label information is converted to the standard input format for HTS engine. The F0 pitch is extracted by a homegrown tool. 39-demensional Mel-frequency cepstrum coefficients (MFCCs) are extracted using the HTK toolkit. All these information (context label information, F0 parameters and MFCCs) compose the final training corpus.

2) Basic unit and state number choosing for HMMs

Cantonese is monosyllabic in nature (like Chinese). Each syllable of Cantonese is often composed of an initial and the final. Coarticulations between adjacent Cantonese syllables are relatively less than inner syllable units (i.e. initial final) [6]. Hence syllable is selected as the primitive of HMM modes. The HMM model generally takes a left-to-right no skipping topological structure for state transitions, and the number of the states is determined by the length of the basic unit. Too little states are not enough to describe the states internal transition, while too many states will increase the unnecessary computational complexity. 9-state HMM models are finally chosen after several comparative experiments. We use the 4-stream Gaussian mixture model (GMM) to describe a syllable pronunciation state model.

3) EM re-estimating and training

The training procedure has a lot of similarity with the speech recognition approaches. The main difference is that both MFCC with their delta and delta-delta dynamic features and logF0 with its dynamic delta features parameters are extracted from corpus and modeled taking into account phonetic, linguistic and prosodic contexts.

4) Context clustering

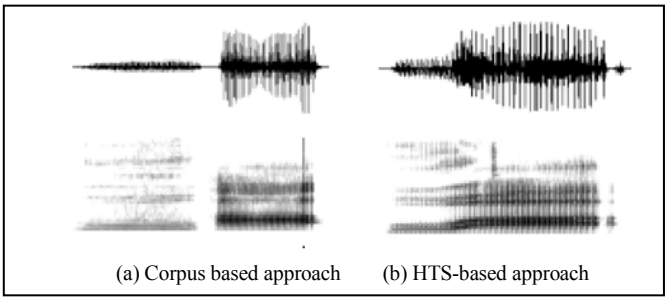
Decision tree based context clustering approach is adopted to cluster the states of trained HMMs, and then the states model parameters in different clusters are shared. Each non-leaf tree node is split into two sub-nodes by a context dependent question like "L-silence?" ("is the previous syllable a silence?") or "C-fa?" ("is the current syllable's jyutping 'fa'?") etc. Leaf nodes of the decision-tree store the distribution of parameters for HMM states, from which HMM parameters for different contexts units can be gained.

B. Synthesis Stage

In the synthesis step, utterance to be synthetized is converted into a sequence of context dependent syllable labels. With this labels, utterance HMMs are constructed. Finally, the waveform speech is synthesized from the gained MFCC and logF0 parameters which are determined using duration distributions with the Mel log spectral approximation(MLSA) filter engine.

III. CONCLUSION

This paper applies the HTS approach to Cantonese speech synthesis by virtue of general architecture of Crystal. Figure 2 compares the differences of the synthetic waveforms (above) and spectrograms (below) between the proposed HTS-based approach and the corpus based concatenative approach. It can be seen that smoother transition in spectrogram can be achieved for the HTS-based approach leading to noticeable improvement in naturalness of the synthetic result for HTS-based approach. Furthermore, the generated synthesis engine of HTS has advantage of small footprint, the size of which is less than 7M bytes, and can be easily ported to embedded electronic devices.



(a) Corpus based approach (b) HTS-based approach

Figure2. The differences of the synthetic result for the word "在/zoi6/" between the HTS-based and corpus based concatenative approaches

Table1. Comparison between HTS and Corpus based Approaches.

	HTS-based	***Corpus based***
Advantage	● Smooth ● Stable ● Small run-time data ● Various voice	● High quality in naturalness
Disadvantage	● Vocoded speech ● Buzzy	● Discontinuity ● Hit or miss ● Large corpus data

REFERENCES

[1] J.Yamagishi. An introduction to HMM-based speech synthesis. Technical report, Tokyo Institute of Technology, October 2006.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis" , Speech Commun., vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[3] Junichi Yamagishi, "An Introduction to HMM-Based Speech Synthesis", 2006

[4] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In Proc. ICASSP-99, pages 229–232, March 1999.

[5] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," in IEEE Speech Synthesis Workshop, 2002.

[6] LAW, K.M. 2001, Cantonese text-to-speech synthesis using sub-syllable units, MPhil. Thesis, Dept. of Electronic Engineering, Chinese University of Hong Kong, 2001.