TRAINABLE CANTONESE/ENGLISH DUAL LANGUAGE SPEECH SYNTHESIS SYSTEM

Haiping Li, Fangxin Chen, Li Qin Shen, Xi Jun Ma

IBM China Research Lab, Beijing 10085, China

ABSTRACT

The Cantonese/English dual language Text To Speech (TTS) system introduced in this paper was developed on IBM's trainable TTS technology, which uses trainable statistical models to automate speech data processing and selection. The Cantonese and English phonological, syntactic and prosodic rules were built into a dual-language Delta module, which processes the mixed-language input accordingly and generates mixed Cantonese and English speech with coherent prosody. To approximate the speaker's characteristics, a speaker prosody profile was extracted from the dataset and incorporated into Delta speech rule processing for the enhancement of duration, lexical tone and intonation prediction. In selection of the concatenative unit set, different Cantonese syllable decomposition schemes were experimented. Though this system is currently only implemented for Cantonese, it can be easily adapted to other tonal languages.

1 INTRODUCTION

In TTS technology, one challenge is how easily a TTS system can generate desired new voices, such as voices for different genders and age groups. This is particularly relevant in commercial applications, where customers always have their own preference. Most existing Mandarin Chinese or Cantonese systems use syllable-based large corpus approach. When constructing such a system, a great amount of continuous speech data is required for one speaker and the phonetic labeling work on the data becomes enormous. As is known, manual labeling of speech data requires phonetic and acoustic expertise. Besides, it is both time-consuming and human-error-prone.

Another challenge in TTS technology is its multi-language ability. It is especially true for a language like Cantonese because of its frequent use of mixed Cantonese and English expressions in speech. A review of existing duel-language TTS systems shows that most duel-language TTS systems support the secondary language in a very limited way. The secondary language inserted into the primary language sounds more like isolated individual words in an alien language environment and not congruous with the primary language's prosody.

Moreover, in a TTS system where the prosody is predicted from the speech rules, the intonation contour, lexical tone and phone duration prediction, etc., may differ significantly from the existing speech data, which could introduce over-modulation to the original speech signal and bring serious voice deterioration to the synthesized speech. How to make the system prosody prediction closer to the existing speech dataset, then, also becomes an important challenge.

The Cantonese TTS system developed at IBM China Research Lab. is based on IBM's trainable speech synthesis system [1]. The advantage of using this approach is that the automatic segmentation and selection of speech units using trainable statistical models becomes possible in handling large speech corpus, and significantly reduces efforts in generate new voices.

New features incorporated into the Cantonese system include methods in dual language intonation prediction, as well as speaker prosody profile build-up to enhance the Front End (FE) prosody prediction matching with the dataset in the Back End (BE). Voice on different syllable decomposition schemes was evaluated for the purpose of optimal concatenation unit selection.

This paper is structured as follows. Section 2 describes the overall system runtime and off-line process; Section 3 gives the description of text normalization process and the prediction of duration and pitch. In Section 4, the speech database preparation and experiments to compare the different phone sets are reported. The conclusions and future work are presented in section 5.

2. SYSTEM DESCRIPTION

The system could functionally be divided into off-line system construction and runtime synthesis; or structurally be divided into FE and BE. FE includes the Romanizer and Delta, which performs the text procession and prosody prediction. BE performs run-time concatenation unit selection and modulation. The whole process of the system is illustrated in Figure 1.

Off-line system construction is for dataset preparation. Dataset consists of context-dependent segment variants selected and indexed from a pre-recorded speech data after signal processing and phonetic unit alignment. For Cantonese dataset, the speech coding included 11 dimensional Mel Frequency Cepstral Coefficients (MFCCs) plus log energy and pitch values. The Speaker Independent (SID) Cantonese HMMs were used to do initial phonetic transcription and silence insertion in the appropriate word boundary. Using the initial alignment, a set of Speaker-Dependent (SD) decision-tree state-cluster HMMs were trained to do final alignment. Based on it, the acoustic decision trees were built for descending to get leaf sequence, and energy trees and duration trees were built for getting target values during runtime synthesis. In order to reduce the size of dataset and improve the runtime speed, pre-selection was performed to keep only the first 25 occurrences of each leaf in the training data. English dataset was constructed in much the same way on the same speaker's English data, except that the log pitch is replaced by one more MFCC coefficient when doing speech coding.

During runtime synthesizing, mixed English and Cantonese text is handled by the Romanizer first. Detailed function of the

Romanizer is described in section 3.1. The output ASCII sequence from Romanizer is fed into the Delta module. This Delta module is to do text normalization, text processing and prosody prediction. The detailed functions of Delta is described in section 3.2. Delta module outputs a sequence of phone segment information, such as pitch, duration, stress, Part Of Speech (POS), and boundary information on syllable, word, and phrase. This sequence is generated on intonation phrase, or language switch bases, and passed to the BE.

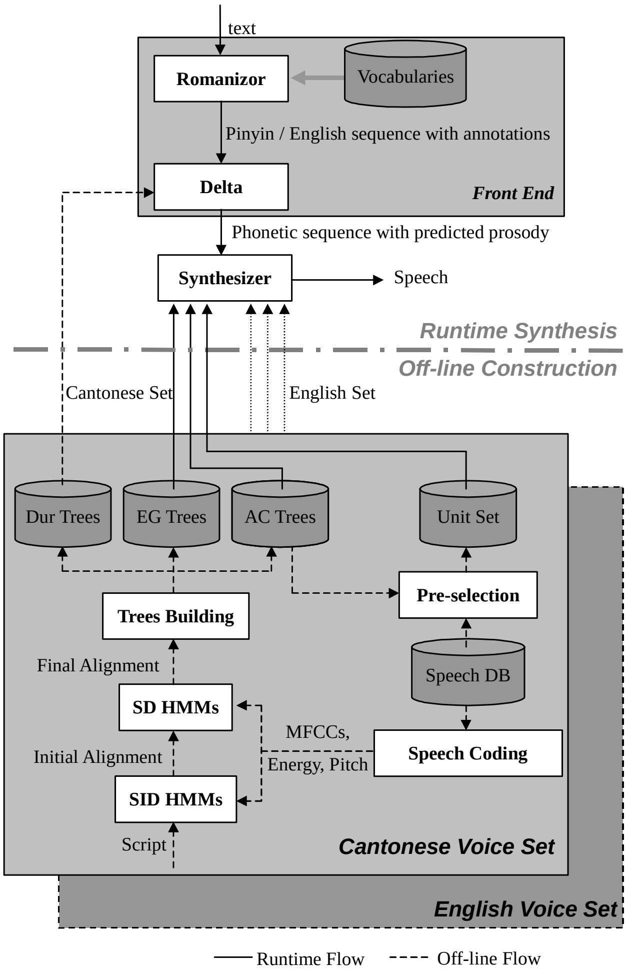

Figure 1: System Process Flow

In the BE synthesizer, Unit selection, modulation and Concatenation is executed independently for Cantonese and English. The phone sequence from FE is first converted into a leaf sequence, by descending the acoustic trees according to the contexts implied by the phone sequences. Target duration for each state are obtained as the median duration of the corresponding acoustic leaf scaled such that the sum of the state durations in each phone is equal to the phone duration specified by the FE. Target F0 values for the end of each state are obtained by linearly interpolating between the points in the F0 contour specified by the FE given the state target duration. Target energy for each state is the median value in each leaf from the energy tree built in the training process. Thus the leaf sequence and target state prosody values are settled down. Then a dynamic programming (d.p.) search is performed over all the waveform segments aligned to each acoustic leaf, with pruning applied in the forward pass to limit the number of parallel segment paths under consideration. Finally, the selected segments are concatenated and modified to have the required prosodic values or the capped prosodic values if applicable.

3. TEXT PROCESSING AND PROSODY PREDICTION

The FE consists of two modules, Romanizer and Delta. The former performs the functions of code conversion, word segmentation, Pinyin conversion, mixed language labeling and POS annotation. The latter is to do text normalization, text processing and prosody prediction.

3.1 Romanizer

The Romanizer module is illustrated in Figure 2.

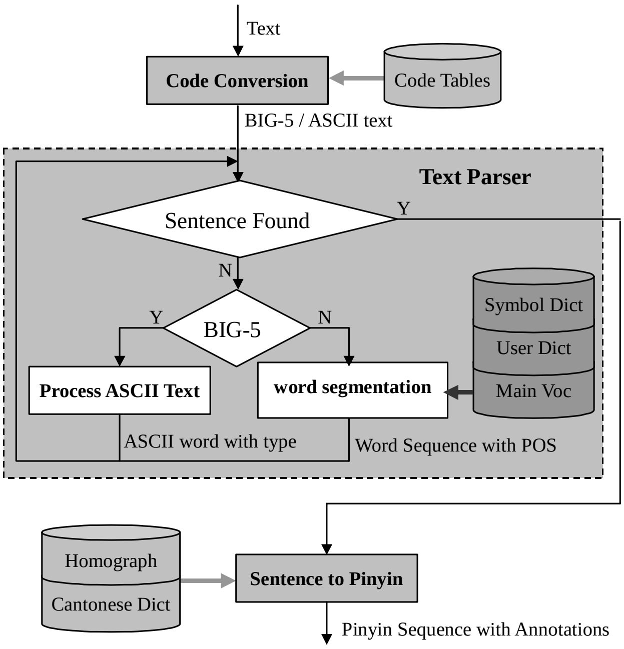

Figure 2: Romanizer Process Flow

When the text is sent to the Romanizer, firstly code conversion is performed. There are different DBCS coding schemes for Cantonese text. The default Cantonese vocabularies implemented in this system are all in BIG-5, therefore, the other coding schemes such as GB and Unicode need to be first converted to BIG-5.

After code conversion, the BIG-5 or ASCII text is fed into the text parser. In this module, text is parsed sentence by sentence. The ASCII text could be foreign words, annotations, space, digits or punctuations, which is picked out and handled first. For the BIG-5 text, word segmentation is performed to find boundaries between Cantonese words. A hybrid word segmentation model is implemented in finding out the word boundary between words, which includes statistical as well as linguistic rules.

The main vocabulary used for word segmentation contains 72 thousand Cantonese words, including the words consisted of the Hong Kong Special Character Set (HK SCS). The POS for each word is also given from this vocabulary after segmentation.

After parsing the input text, pinyin is generated in the sentence-to-pinyin procedure. To get the pinyin for the words in the sentence, first the homograph dictionary is searched. The Cantonese homograph dictionary contains 9 thousand words. If the word is not in the dictionary, then search a set of Cantonese characters for applying special homograph rules. If the target word is still not in the list, then the default pinyin rules will be applied to it.

The final Romanizer output to Delta is a pinyin sequence with word boundaries, as well as annotations for text properties such as POS, stress patterns, and language type.

3.2 Dual-language Delta Module

Figure 3 illustrates the process flow chart of the dual-language Delta module. Delta is fundamentally a rule-based text process system, where the language-universal as well as language-specific phonetic, syntactic and acoustic rules are applied to the input text. Text Normalization handles the specific rules in reading text, such as how to read digits, dates, abbreviations, web sites, etc. Text Processing applies various language-related rules to the phonetic representation of linguistic units at different levels. Prosody Prediction applies acoustic rules in pitch, duration, energy and pause. As shown in Figure 3, the dual-language Delta module performs text normalization, text processing separately for Cantonese and English, but treated them together in prosody prediction, which ensures a coherent intonation in mixed language situation.

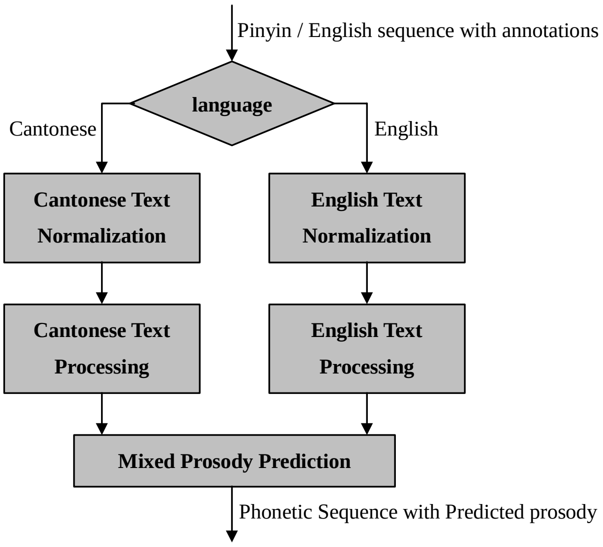
Pinyin / English sequence with annotations
language
Cantonese
English
Cantonese Text Normalization
English Text Normalization
Cantonese Text Processing
English Text Processing
Mixed Prosody Prediction
Phonetic Sequence with Predicted prosody

Figure 3: Delta Architecture

3.2.1 Intonation Prediction

The complication of intonation prediction for a Cantonese-English dual-language system is in two folds: First, Cantonese is a contour tonal language. It has nine distinctive lexical tones. The pitch contour of Cantonese tone is at the syllable level and it is superimposed on the intonation at the sentence level. This adds one additional dimension in pitch prediction as compared with non-tonal languages. Secondly, in the mixed language situation, the intonation can not be handled separately as in their individual languages. Observation of mixed language speech shows that the secondary language words are mostly merged into the primary language's global intonation, but still keep their own pitch patterns in a relative manner.

The intonation prediction algorithm implemented in this system is similar to Fujisaki's intonation model, which consists of two parts: the intonation phrase part and tonal/accent part. The intonation phrase part consists of an exponential F0 decay function. This part is shared by both Cantonese and English, which ensures a coherent intonation pattern in the mixed language situation. The tonal/accent part is language dependent. In the Cantonese situation, The Chao's five-level tone scheme is implemented to define the target pitch levels for different lexical tones [2]. The target pitch level is also subject to tonal co-articulation rules as well as other syntactic rules. In the English situation, the pitch target is based on pitch accent, phrase tone and boundary tones. In both Cantonese and English situations, the pitch values determined from tonal/accent part are superimposed onto the pitch values determined from the intonation phrase.

Since the general intonation model is speaker-independent, a speaker prosody profile is required for each existing dataset to set the necessary values such as speaker's pitch range, pitch decay degree, and pitch patterns for different lexical tones. The speaker prosody profile is offline extracted from the dataset, which included not only the pitch information, but also phone duration information which will be discussed later. A statistical example for phone AA in tone 1 extracted from dataset is illustrated in figure 4. The phrase pattern considered is a sentence consisting of main and subordinate clauses.

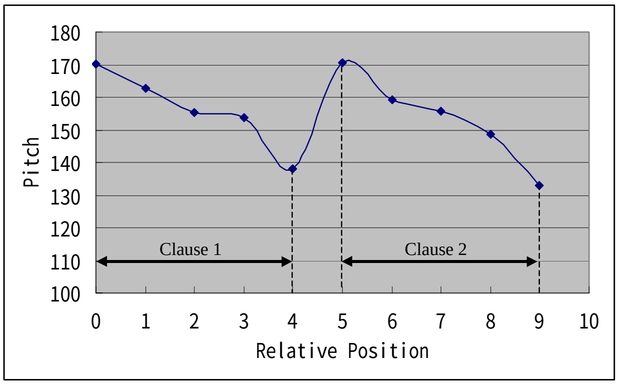

Figure 4: Relationship of pitch value and position

The mean F0 values at different positions of both clauses initial and final positions for the phone AA1 are to be referenced by the general intonation model in pitch prediction. This measure makes the predicted target values closer to unit selected by the BE and reduces the possible distortion introduced by signal over-modulation. The current prosody prediction is a state-of-the-art work. The accumulated tuning is necessary to handle variations and special cases.

3.2.2 Duration Prediction

Two different duration prediction mechanisms were implemented in the current Cantonese system. One is a hybrid approach, which uses both rule and statistics. In this approach, the initial phone durations in different phonetic context are extracted from the BE voice set and stored in the speaker prosody profile. Delta sets those phone values as default and further processes them with various speech rules such as phone duration in different syntactic position, stress level, and POS.

Another duration prediction approach is using the duration trees built on the BE training data. Such trees are constructed automatically to maximize the leaf distance. If this option is selected, the FE phone duration prediction will be ignored and the duration trees in the BE is solely responsible for phone duration selection.

The using of speech rules could compensate the speech data sparse problem, while using the duration trees makes the TTS system more trainable. Both have their own advantages. Since the current Cantonese dataset is trained only on about 2 thousand sentences and the data did not cover many important phone contexts which could influence the phone durations. Therefore, the default setting for duration prediction is using the hybrid method. But researchers still have the option of using duration trees.

4. OFF-LINE SYSTEM CONSTRUCTION

The main steps of the system construction are described in section 2. Only the speech database preparation and experiments to compare the different phone sets are reported here.

4.1 Speech Database Preparation

The script for recording the Cantonese speech database was generated by running a Greedy Algorithm over a Cantonese text Corpus consisting of about 35 thousand sentences collected in newspaper domain. In this algorithm, score for each sentence was based on the number of different phones in different phonetic contexts. The phones were defined by using the main vowel Cantonese syllable decomposition method [3]. And the phone variations were the phones in different tone and phone contexts [4]. The process was based on the pinyin output of the Corpus generated by the FE. After selection, the script used in this Cantonese TTS system included 2201 sentences with the length varying from 10 to 30 syllables.

For English speech database, we used the same script designed in [1]. The two speech databases were all recorded at 22kHz sample rate, with one channel for speech and the other for laryngograph signal. After recording, the consistency of the recorded speech with the script was manually checked.

4.2 Experiments on Different Phone Sets

One key issue in constructing this Cantonese TTS system was designing the phone set for building the HMMs. For Cantonese, there are different schemes to decompose syllables. The two schemes (denoted here as A and B) described in [3] were used to generate the phone sets for experiments. In scheme A, a syllable was decomposed into *onset*, *nucleus* and *coda.* And the tonal information was supposed to be carried only on the nucleus. Thus the codas from syllables with different tone were put into one acoustic tree. In scheme B, a syllable was decomposed into *initial* and *final*. The lexical tone was contained in the final.

To compare the synthesis quality, two systems using different phone sets were constructed. Five pairs of synthesized samples from these two systems were evaluated by voting which sample was better. These sentences were out of the training database. The order in each pair was randomly set. Thus the 5 evaluators did not know which sample was from which system. The results are shown in table 4. The two schemes did not have much difference in performance. In this Cantonese Synthesis System the Scheme B was implemented.

Samples	Vote A	Same	Vote B
Pair 1	3	1	1
Pair 2	0	4	1
Pair 3	2	2	1
Pair 4	1	2	2
Pair 5	1	2	3
Sum	**7**	**11**	**7**

Table 4: Vote Number of Systems using Different Size of Units

5. CONCLUSIONS AND FUTURE WORK

Currently this IBM Cantonese TTS system is being developed into product, and also integrated into the Cantonese Websphere Voice Server (WVS) 2.0. The demonstration using this synthesis system is available by calling 8610-62986677-257.

Improving work under way includes exploring new methods for script design, word segmentation and prosodic unit grouping based on statistical methods. Experiments have been conducted and the initial results are promising.

6. REFERENCE

[1] Donovan, R.E., Eide, E.M., “The IBM Trainable Speech Synthesis System”, *Proc. ICSLP’98*, Sydney

[2] Fangxin Chen, “Issues in Speech Synthesis for Tonal Languages”, *Proc. SNLP-oriental COCOSDA 2002*, Thailand

[3] C. J. Chen, Haiping Li, Liqin Shen, Guokang Fu, “Recognize Tone Languages using Pitch Information on the Main Vowel of Each Syllable”, *Proc. ICASSP2001*, Salt Lake City

[4] Haiping Li, Fangxin Chen, Liqin Shen, “Generating Script Using Statistical Information of the Context Variation Unit Vector”, *Proc. ICSLP2002*, Denver.