CU VOCAL: CORPUS-BASED SYLLABLE CONCATENATION FOR CHINESE SPEECH SYNTHESIS ACROSS DOMAINS AND DIALECTS

Helen M. Meng, Chi Kin Keung*, Kai Chung Siu*, Tien Ying Fung* and P. C. Ching***

*Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
**Digital Signal Processing Laboratory,
Department of Electronic Engineering,
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
{hmmeng,ckkeung,siukc,tyfung@se.cuhk.edu.hk, pcching@ee.cuhk.edu.hk}

ABSTRACT

This paper describes CU VOCAL, a Chinese text-to-speech synthesis system that adopts the approach of corpus-based syllable concatenation. We have demonstrated the applicability of the approach primarily for Cantonese, a major dialect of Chinese predominant in Hong Kong, South China and many overseas Chinese communities. This work extends our previous work as described in [1]. Our approach is able to synthesize speech from free-form text, and it can also be optimized for response generation in specific application domains. We have also demonstrated the portability of the approach to Putonghua, the official Chinese dialect, in a domain-optimized setting. Coarticulatory context is expressed in terms of distinctive features. Tonal context is also included. We conducted a series of listening tests using CU VOCAL, which gave favorable performance.

1. INTRODUCTION

This paper reports on our recent work in applying corpus-based syllable concatenation for Chinese text-to-speech synthesis. We focus on Cantonese, a major Chinese dialect predominant in Hong Kong, South China and many overseas Chinese communities. The corpus-based concatenation technique has been gaining popularity in speech synthesis [2-6] due to its ability to achieve a high degree of naturalness. The use of corpus-based syllable concatenation is particularly suitable for Chinese, since the language is monosyllabic in nature. A compact inventory of tonal syllables provides complete phonological coverage for a given dialect. For example, Cantonese has 20 syllable initials and 53 syllable finals, constituting 660 base syllables. The dialect also has 6 lexical tones, giving a total of about 1800 tonal syllables. For Putonghua, the official dialect of Chinese, there are 24 syllable initials and 37 syllable finals, constituting about 410 based syllables. Putonghua also has 5 lexical tones that form about 1,400 tonal syllables. Hence, similar to other work in Chinese concatenative synthesis, e.g. [6,7], we use the tonal syllable as the basic unit for concatenation. As pointed out in [8], the fundamental frequency corresponding to the tone of a given (tonal) syllable may vary due to differences in tonal context. Hence when we select syllable units for concatenation, we take into consideration the tones of the neighboring syllables. Co-articulatory effects are also considered in terms of distinctive features. Distinctive features are minimal linguistic units that distinguish between speech sounds, e.g. LABIAL refers to using the lips and VELAR refers to raising the velum that separates the nasal and oral cavities. It is believed that about twenty or so features can characterize all the languages in the world. In short, our concatenative synthesis approach selects tonal syllables by considering tonal context and place of articulation in terms of distinctive features.

In the following, we will describe the various components and techniques used in the CU VOCAL system, as well as present results from a series of listening tests.

2. TEXT NORMALIZATION

Text normalization is a critical step in Chinese text-to-speech synthesis. This step ensures that (i) the sequence of Chinese characters in the input text is tokenized properly into a word sequence; (ii) concepts are verbalized appropriately; (iii) named entities is identified correctly; and (iv) mixed language can be suitably handled, especially for the linguistic environment in Hong Kong.

The Chinese word may consist of one to several characters and there is no explicit word delimiter. Word tokenization contains much ambiguity. However, locating the word boundaries correctly is very important for pause insertion and correct pronunciation lookup. Most segmentation algorithms are dictionary-based which require a pre-compiled word list or lexicon. [9] We have developed a Chinese lexicon with over 200,000 entries based on [10]. Word tokenization uses a maximum matching algorithm with reference to the lexicon. This matching algorithm may proceed from left to right (i.e. forward match) or from right to left (i.e. backward match). Segmentations from the forward and backward matches are compared, and discrepancies are resolved according to heuristics that favor fewer segments and longer segments. An example is shown in Table 1.

It is also possible that the resulting segmentation is a combination of the forward and backward matches. A pause is inserted at the word boundaries after every few words. Our procedure also looks up the pronunciations of the tokenized words from the 200,000 pronunciation dictionary. Many Chinese characters have multiple syllable pronunciations, depending on the lexical context. Hence segmenting correct word boundaries is very important for mapping to the correct pronunciation. For example, the character 仔 is often pronounced as /zai2/, except for the word 仔細 /zi2-sai3/ .

Original text input:
大約有 ５００ 名自稱爲學生的激進分子
(approximate translation: *approximately five hundred extremists who claim to be students*)

Convert entirely into characters:
大約有五零零名自稱爲學生的激進分子

Forward match (left-to-right tokenization):
大約有 五零零 名 自稱爲 學生 的 激進分子
^^^^^^^^^^^^^^
(approximate translation: approximately five zero zero extremists who claim to be students)

Backward match (right-to-left tokenization):
大約有 五 零零 名 自稱 爲學 生的 激進分子
^^^^^^^^^^^^^^
(this word sequence cannot be translated properly)

Selected tokenization using heuristics:
forward match

Further text normalization: (e.g. numeric expression)
大約有 五百 名 自稱爲 學生 的 激進分子
(convert "five zero zero" to "five hundred" since it is identified to be a count via heuristics)

Table 1. Example illustrating text normalization used in CU VOCAL.

We have also written a set of heuristics to verbalize special concepts appropriately, such as for dates (e.g. 10/1/2001 should not be pronounced as “一零一二零零一”, but rather “二零零一年十月一日”), times (e.g. 7:30pm should not be pronounced as “七三零 PM” but rather “下午七時三十分”) and numeric expressions (e.g. "500" was corrected from “五零零” to “五百” in Table 1 above).

Named entities also require special handling. Our Chinese text normalization can automatically tag the names of people by referring to the 500 most common Chinese last names. These have characters that are also commonly used as normal words or part of a normal word, but the same character is often pronounced differently when it is a Chinese surname. For example, 單 is often pronounced as /daan1/ (e.g. 單獨 /daan1-duk6/, i.e. “alone”) but becomes /sin6/ when the character is used as the last name of a person.

Our text normalization can also handle mixed language (between Chinese and English) especially since newspaper text in Hong Kong often include URLs, email addresses, English acronyms and words. URLs, emails and English acronyms are identified by a set of heuristic rules. They are synthesized by reading the sequence of alphabetic letters and punctuations. Other English text are handled separately by an off-the-shelf English speech synthesizer. We use the FESTIVAL[11] English text-to-speech synthesis system for this feature.

3. SPEECH DATABASE DEVELOPMENT

We need to develop a speech database that is compact but also achieves high coverage of the syllable units and their contextual variants that will be needed in concatenative synthesis. To this end, we collected a large corpus of Chinese text from a diversity of sources, and covering a variety of topics. The corpus contains over half a million sentences which we segmented and converted into their tonal syllable pronunciations based on pronunciation lookup. The representation of each tonal syllable pronunciation is also augmented with four contextual features comprising places of articulation (i.e. distinctive features) [1] and tones of its left and right neighbors. We use the filtering algorithm illustrated in Table 2 to select the minimum number of sentences that provides maximum coverage of the syllable contextual variants.

Step 1: Compile the set of distinct syllable units in the corpus.
Step 2: Compute a score for each unit
score = 1/no. of occurrences
Step 3: Calculate the score of each sentence
Sentence score = Σ acoustic unit scores
If all sentences score zero, END.
Step 4: Sort the sentences in by their scores. Move the highest scoring sentence from the corpus into the *filtered set*.
Step 5: Reset all scores in the *reduced* corpus to zero. Goto Step 1.

Table 2. Algorithm for selecting sentences from a large corpus into a small set of recording prompts which maximizes phonological coverage of Chinese tonal syllables and their contextual variants.

The algorithm filtered our corpus down to approximately 600 sentences that contain about 1500 unique tonal syllables. We augment this set with another 330 tonal syllable segments from CU SYL [12] – a syllable corpus developed in-house. These sentences are recorded and the sentence waveforms are segmented into syllable speech segments by forced alignment with an HMM-based syllable recognizer. The speech segment boundaries are then verified manually.

4. UNIT SELECTION

Given some input text, CU VOCAL invokes the text normalization component for word tokenization, verbalization and conversion into a sequence of tonal syllables. Desired contextual features (distinctive features and tones) for each syllable unit are also derived in this step. Concatenation then proceeds from left to right, and unit selection aims to pick from our speech database the syllable unit that provides the best match in terms of the desired contextual features. Largest syllable units in speech database are selected first. Thereafter, we select

among the contextual variants for each tonal syllable by minimizing the cost as defined in Equation (1):

$$Cost = \sum_i w_i Dist(df_i, af_i).....(1)$$

where
df_i is the *ith* desired feature for a given syllable unit;
af_i is the *ith* available feature for the same syllable unit in the speech database;
Dist denotes the distance between the two features; and
w_i is a weight that indicates the relative importance of feature *i*.

We have performed our unit selection based on two kinds of contextual features. The first is place of articulation, which is described in [1]. The second is tone. Tonal context is important for Chinese synthesis as Chinese is a tonal language.

Values for w_i and $Dist(df_i, af_i)$ that correspond to distinctive features are manually assigned, with reference to phonological theory and results from perceptual tests. We have also developed the following scheme for unit selection based on tonal features. Detailed justifications based on listening tests are presented in [13]. According to the scheme, the ideal tonal variant is one with matching left and right tonal context. If the ideal case cannot be found, we enforce a match in left tonal context; otherwise, we follow the incremental matching rule described as follows:

1. We favor the syllable unit that maintains the slope in the tone trajectory going from the preceding syllable unit to the current syllable unit.
2. If condition (1) is satisfied, we will try to find the syllable unit whose left tonal context has the same tone shape as that of the ideal syllable unit.
3. If conditions (1) and (2) are satisfied, we will try to find the syllable unit that minimized transitional movements in the tone trajectory going from the preceding syllable unit to the current syllable unit.
4. We avoid using a syllable unit that has tone 2 as its left tonal context due to the dynamic and transitional nature of the tone.

5. DOMAIN OPTIMIZATION AND PORTABILITY ACROSS DIALECTS

5.1 Domain Optimization

Our approach is amenable to optimization to specific domains. Optimization involves enhancements to the speech database based on domain-specific knowledge, with the aim to minimize unit concatenation costs (to be described later) to improve the quality in synthesis outputs. We have worked on two broad domains:

- press releases from the Hong Kong SAR government;
- financial news;

and three constrained domains:

- air travel planning;
- real-time stock quotes; and
- real-time foreign exchange information.

Domain optimization typically begins with the collection of domain-specific text data. From these we extract terminologies that frequently occur within the domain. A special vocabulary list will be created containing extracted terminologies absent from the existing (domain-independent) speech database. To augment the speech database for the specific domain, we select sentences from the domain-specific text data based on the special vocabulary list to generate the additional recording prompts.

This procedure is designed to maximize re-use of syllables from the existing speech database and minimize the effort required in additional recording and syllable segmentation. For example, for the financial news domain, the additional recordings include approximately 170 sentences covering about 360 terms.

5.2 Portability Across Chinese Dialects

We have also investigated the portability of our approach across Chinese dialects for domain-specific applications. We replaced the tones and distinctive features for Cantonese with those for Putonghua. The remaining part of our methodology remains the same. We have developed a constrained Putonghua speech database for the stocks and foreign exchange domains. This component has been integrated into a trilingual spoken dialog system known as ISIS [14].

6. EVALUATION

6.1 Listening Comprehension

In order to evaluate the quality of the CU VOCAL system for domain-independent synthesis, we selected a news article for synthesis. The generated waveform has a duration of about one minute, and was played to 10 subjects. The subjects have been asked to listen attentively and are free to take notes if they wish. After listening they are given a set of 5 questions about the news story, and are asked to write their answers down for scoring. The news story and the questions are designed such that the listeners cannot use their general knowledge to answer the questions, but rather they have to rely entirely on what they have heard during the listening comprehension task. The story and the questions are shown in Tables 3 and 4 respectively.

在地產市道持續低迷下，規劃環境地政局及行政署已初步商定維持原議把添馬艦地王用作興建政府總部，新總部大樓預計在二零零七至零八年間落成啓用，政府預期整項計劃可以創造最少五千個就業機會，並在明年初先進行設計比賽爲私人企業製造商機，不過，政府高層認爲興建政府總部雖有迫切性，但在現時經濟低迷時大興土木是否獲得市民支持，卻感到猶疑，事件短期內會交行政會議最後拍板。

Table 3. News text used in synthesis for listening comprehension test.

1.規劃環境地政局及那一個機構已初步商定維持原議把添馬艦地王用作興建政府總部？ 2.新總部大樓預計在那年落成啓用？ 3.政府預期整項計劃可以創造最少幾多個就業機會？ 4.在明年初舉行什麼活動爲私人企業製造商機？ 5.政府高層擔心在現時經濟低迷時大興土木未必能獲得那一方面的支持？

Table 4. Questions in the listening comprehension test.

Every question is scored with 0 point for errors, 1 point for a partially correct answer and 2 points for a correct answer. On average our subjects obtain 7.7 out of 10 points. They have

also been asked to rate the naturalness of the synthesis output on the scale of 1 (least natural) to 6 (most natural, equivalent to having the passage read by a human). CU VOCAL obtained a mean opinion score of 3.4 from our 10 subjects. For the sake of comparison, a similar test using a PSOLA-based synthesizer gave a mean opinion score of 2.8.

6.2 The Effect of Domain Optimization

We attempt to assess the effect of domain optimization using a listening test. 10 sentences were selected – 3 from the foreign exchange domain, 3 from the stocks domain and 4 from the air travel domain. Each sentence is synthesized twice to give a pair of waveforms – one using CU VOCAL with domain optimization, the other without. The 10 pairs of waveforms were played to the 10 subjects who were again asked to rate the naturalness of each waveform on the scale of 1 (least natural) to 6 (most natural). The order in each pair of waveforms has been randomized. The mean opinion score without domain optimization was 3.1, which interestingly, was very close to that in the listening comprehension test. The mean opinion score with domain optimization rose to 4.7. The increment was statistically significant, based on a paired *t*-test with α=0.05. These results show that the procedure of domain-optimization can effectively enhance the naturalness of the synthesized speech.

7. CONCLUSIONS AND FUTURE WORK

This paper presents the CU VOCAL system, a Cantonese concatenative speech synthesis system. The tonal syllable is used as the basic unit for synthesis, and coarticulatory context is captured in terms of tonal features and distinctive features. We present methods used in text processing, in speech database development and unit selection to minimize costs due to mismatched coarticulatory features. CU VOCAL is amenable to optimization for specific domains, and its synthesis methodology is also portable from Cantonese to Putonghua simply by altering the set of tonal features and distinctive features. Evaluation based on a listening comprehension task and a listening test indicates that the synthesized output from CU VOCAL is highly intelligible and reasonably natural. The naturalness can also be significantly enhanced by the domain optimization procedure proposed in this paper.

8. ACKNOWLEDGMENTS

This project is funded by a Direct Grant from The Chinese University of Hong Kong.

9. REFERENCES

1. Fung, T. Y. and H. Meng, "Concatenating Syllables for Response Generation in Spoken Language Applications," Proceedings of ICASSP 2000.
2. Campbell, N. and A. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis," Progress in Speech Synthesis, Springer Verlag, 1996, pp. 279-282.
3. Hon, H. et al., "Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems," Proceedings of ICASSP 1998.
4. Donovan, R. E. et al., "Phrase Splicing and Variable Substitution using the IBM Trainable Speech Synthesis System," Proceedings of ICASSP 1999.
5. Yi, J. and J. Glass, "Natural Sounding Speech Synthesis using Variable-Length Units," Proceedings of Eurospeech 1999.
6. Chu, M., et al., "Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer," Proceedings of ICASSP 2001.
7. Lee, L. S., C. Y. Tseng and M. Ouh-Young, "The Synthesis Rules in a Chinese Text-to-Speech," IEEE Transactions on Acoustics, Speech and Signal Processing, 37(9), pp. 1309-1320, 1989.
8. Shih, C. L. and G. P. Kochanshi, "Chinese Tone Modeling with Stem-ML," Proceedings of ICSLP 2000.
9. Sproat, R. et al., "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," Computational Linguistics, 22(3), 1996.
10. CULEX. http://dsp.ee.cuhk.edu.hk/speech, 1999.
11. Taylor, P. et al., "The Architecture of the Festival Speech Synthesis System," Proceedings of the third ESCA Workshop on Speech Synthesis, pp. 147-151.
12. Lee, T., W. K. Lo, P. C. Ching and H. Meng, "Spoken Language Resources for Cantonese Speech Processing," Speech Communication, forthcoming.
13. Fung, T. Y. and H. Meng, "A Study of the Effect of Tonal Context on Chinese Concatenative Speech Synthesis," submitted to the International Symposium on Chinese Spoken Language Processing, 2002.
14. Meng et al., "ISIS:A Learning System with Combined Interaction and Delegation Dialogs" Proceedings of Eurospeech, 2001.