Concatenating Syllables for Response Generation in Spoken Language Applications

Tien Ying Fung and Helen M. Meng
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China
{tyfung@se.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk}

ABSTRACT

We describe our approach in developing a speech synthesis technique for response generation in domain-specific spoken language applications. Our approach handles two Chinese dialects – Cantonese and Putonghua. We chose the foreign exchange domain, and worked with its constrained vocabulary and response expressions. The syllable is selected to be our basic unit for concatenation. Each unit label includes a two-digit appendix to encode the distinctive features of the left and right coarticulatory context. Our approach attempts to maximize intelligibility and naturalness of the generated responses within the application domain. Hence the synthesized outputs compare favorably with a domain-independent TD-PSOLA synthesizer.

1. INTRODUCTION

Recently we see a surge in the development of spoken language systems that enable the computer to answer the user's informational queries regarding a restricted domain. A critical factor contributing towards the habitability of such a human-computer speech interface is a high degree of naturalness and intelligibility in the spoken presentation of relevant information (i.e. the generated response).

A key technology for response generation is speech synthesis. In this area, the use of corpus-based concatenation techniques has been gaining popularity [1-6], due to their ability to achieve a high degree of naturalness. This paper reports on our initial attempt in designing a process for generating speech in two Chinese dialects, by means of concatenating syllables. The generated speech serves as the response of a spoken language system in the financial information domain. We chose the syllable to be our basic unit for concatenation, since the Chinese language is monosyllabic in nature. However, we also include multi-syllable units, derived from the frequently occurring words and phrases in the application domain. This should minimize the number of concatenations needed and consequently the perceived distortions they bring to the generated waveforms.

2. BACKGROUND – APPLICATION DOMAIN

Our application domain is well-suited for Hong Kong. The region has a trilingual populace speaking Cantonese, Mandarin and English. It is also one of the major finanical centers in the world. Both landline and mobile phone penetrations are high – the former is near saturation, and the latter is over 50%. Our long term goal is to develop a trilingual human-computer speech interface (complete with speech recognition and speech synthesis), to enable users to access real-time financial information via speech. As an initial step, we work within the foreign exchange domain, and capture real-time data through a dedicated Reuters satellite feed [7,8]. In this domain, generated responses need to provide information about dates, times, currencies to buy / sell, bid / ask prices in exchange rates, etc.

3. OUR APPROACH

Our response generation component performs the tasks of text generation (for both English and Chinese), followed by speech synthesis. A grammar is written for the various information categories. The grammar can run in generative mode, and combine with the raw data from our satellite feed, to produce a verbalized presentation of the information. We send our generated English text to the FESTIVAL system [9] for synthesis of English. FESTIVAL has been made freely available for research by the University of Edinburgh. Chinese text is sent to our syllable concatenation module to generate Putonghua

and Cantonese. The development of this module consists of the following steps:

1. Corpus development for recording
2. Waveform segmentation
3. Unit selection for concatenative synthesis

4. CORPUS DEVELOPMENT

Corpus development is a important step in our approach. We aim to design a set of recording prompts that fully cover the vocabulary of the domain. The recording prompts also need to include different realizations of an acoustic unit due to coarticulatory effects or variations in prosodic contexts. This coverage should hopefully be achieved with the minimum number of recording prompts. Therefore our approach aims for a high degree of intelligibility and naturalness within the scope of an application domain. It has greater flexibility than using entire pre-recorded responses for an application, but it is more restrictive than a full TTS system which is not bound by the scope of the domain.

The tonal syllable is adopted as our basic acoustic unit for synthesis. Cantonese has six tones and Putonghua has five tones. Frequently occurring words (e.g. currency names) are treated as a single multi-syllable unit. Each unit has a two-digit appendix to encode its left and right coarticulatory context.

4.1 Generate-and-Filter

We use a "generate-and-filter" algorithm to produce a compact set of recording prompts. First the response grammar is used to generate the possible response expressions. We refer to this as our *generated set*. It includes carrier phrases with the appropriate prosodics for our application domain. These sentences are transformed into syllable and multi-syllable units by looking up their pronunciations from dictionaries. The filtering process is illustrated in Figure 1. It aims to compress the *generated set* but retain all the contextual variations of the existing acoustic units. This compressed set becomes our *filtered set.*

Table 1 shows the size of our generated and filtered set of sentences. We see that our generate-and-filter algorithm is able to compress roughly by a factor of 8. The resultant *filtered set* is of a manageable size for recording.

Step 1: Compile the set of distinct acoustic units (with the two-digit context encoding) from the *generated set*.

Step 2: Compute a score for each acoustic 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 *generated set* into the *filtered set.*

Step 5: Reset all scores in the *generated set* to zero. Goto Step 1.

Figure 1. Algorithmic flow of our filtering mechanism for producing a compact set of recording prompts.

Language	Generated Set	Filtered Set
Cantonese	3860 sentences	450 sentences
Putonghua	4870 sentences	650 sentences

Table 1. No. of sentences in the generated set and filtered set respectively. Our generate-and-filter algorithm compresses the generated set by approximately 8 times, to produce the recording prompts.

4.2 Coarticulation and the Use of Distinctive Features

It is widely known that coarticulatory effects from the left and right contexts can substantially change the acoustic realization of an acoustic unit. For example, consider the character 七 (i.e. the number '7'), which is pronounced as 'cat1' in Cantonese.[1] In the context of "六七八"(i.e. the number sequence '678', pronounced as "luk6 cat1 baat3"), the syllable 'cat1' has a *left velar* context, and a *right labial*

[1] Our Cantonese transcriptions follow the LSHK standard, set by the Lingustic Society of Hong Kong [10].

context. Due to coarticulation, speakers tend to assimilate the alveolar closure of the syllable 'cat1' with the right labial, resulting in the production of 'cap1' (e.g. 輯). Hence if we were to extract the syllable acoustic wave file for 七 from the spoken phrase "六七八", and use it to synthesize "八七六", the resulting waveform will sound like "八輯六", which will be perceived as incorrect.

In this work we append a digit pair to the labels of every acoustic unit. The digits encode the place of articulation of the (optional) coda of the left syllable, and the (optional) onset of the right syllable. The distinctive features that are considered for Cantonese and Putonghua are tabulated in Tables 2 and 3 respectively. Incorporating additional contextual information should improve the quality of synthesis, but also increase the size of the acoustic wave repository.

CANTONESE	
Right Context	*Left Context*
Alveolar	Alveolar
Glide	Labial
Neutral	Velar
Labial	--
Lateral	--
Palatal	--
Velar	--

Table 2. Distinctive features used to represent the left and right context for Cantonese acoustic units.

PUTONGHUA	
Right Context	*Left Context*
Alveolar	Alveolar
Dental-Alveolar	Back
Glide	Glide
Labial	Mid-Front
Lateral	Retroflex
Neutral	Rounded
Palatal	Velar
Retroflex	--
Velar	--

Table 3. Distinctive features used to represent the left and right context for Putonghua acoustic units.

5. WAVEFORM SEGMENTATION

The segmentation of the recorded sentences is carried out manually using spectrograms. We have also begun to use a syllable recognizer to provide a forced alignment as an intial segmentation, to be hand-refined as a next step. Figure 2 shows a spectrogram of the Cantonese sentence 七千八百三十一點八三一 (translation: seven thousand eight hundred and thirty one point eight three one). The segmentation process is also important for the quality of the synthesis outputs. We recorded from two female speakers, one for Cantonese and the other for Putonghua. The process of segmentation yields 2400 acoustic segments.

6. UNIT SELECTION FOR CONCATENATIVE SYNTHESIS

Henceforth our unit selection process is simple. For a given textual input which is mapped into syllable-based units, our synthesis algorithm concatenates the corresponding acoustic wave files sequentially from left to right. The unit selection process ensures that the acoustic segments with matching left and right contexts are chosen. We also inserted short pauses in between phrases, and long pauses in between sentences. Both the unit selection process and the insertion of pauses were found to be important contributing factors towards naturalness in the synthesized outputs.

7. A LISTENING TEST

We conducted a listening test to evaluate the quality of our synthesized responses. Our benchmark is a full Cantonese TTS synthesizer developed internally [11]. This synthesizer can handle domain-independent textual input by TD-PSOLA synthesis. In comparison, our current synthesis task is simpler and more restrictive due to domain-specificity. We hope to show that the effort devoted to optimization within the domain contributes towards higher intelligibility and naturalness of the synthesized outputs. On the other hand, the TD-PSOLA synthesizer has no optimization with respect to the application domain at all.

The listening test is set up as a within-group experiment involving 12 subjects. Ten pairs of synthesis outputs are generated from ten sentences,

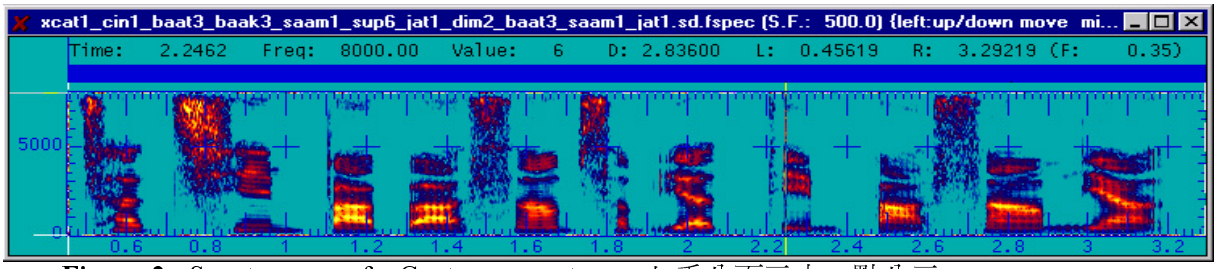

Figure 2. Spectrogram of a Cantonese sentence, 七千八百三十一點八三一

i.e. two waveforms per sentence. The sentences cover all the currencies within our foreign exchange domain, and their exchange rates at various dates and times. One of the waveform pairs is generated by the TD-PSOLA synthesizer, and the other by the current syllable concatenation technique. The order of the waveforms are randomized to neutralize learning effects. Each subject is asked to rate the pair of waveforms in terms of intelligibility and naturalness, on a scale of *1* to *6* (*1* represents barely intelligible / natural, and *6* represents extremely intelligible / natural).

We formulated a *t-test* using the difference in opinion score as our test statistic. The differences in intelligibility scores have a mean of 1.2 and a standard deviation of 1.17. The differences in naturalness scores have a mean of 1.8 and a standard deviation of 1.15. Testing at a significance level of 0.05 concludes that we should accept the alternate hypothesis. i.e. syllable concatenation is more intelligible and natural than TD-PSOLA within the foreign exchange application.

8. CONCLUSIONS AND FUTURE WORK

This paper reports on our approach in developing a speech synthesis technique for Chinese response generation in a domain-specific spoken language application. Our approach handles two Chinese dialects – Cantonese and Mandarin. We chose the foreign exchange domain, and worked with its constrained vocabulary and response expressions. The syllable is selected to be our basic unit for concatenation. Each unit label includes a two-digit appendix to encode the distinctive features of the left and right coarticulatory context. Our approach aims to maximize the intelligibility and naturalness of generated responses within the scope of the domain. Hence the synthesized outputs compare favorably with a domain-independent TD-PSOLA synthesizer.

In the near future, we plan to include tone considerations in our synthesis process, as well as extend the scope of our application domain.

ACKNOWLEDGMENTS

The authors wish to thank Tan Lee of the Digital Speech Processing Laboratory for assistance with the listening test setup. The second author also thanks her undergraduate project students – Brenda Chan, Jessica Hui and Angel Lau for various implementational assistance.

REFERENCES

[1] 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.

[2] Donovan, R.E. et al., "Phrase Splicing and Variable Substitution using the IBM Trainable Speech Synthesis System," Proceedings of *ICASSP*, 1999.

[3] Chou F. C. et al., "Selection of Waveform Units for Corpus-based Mandarin Speech Synthesis based on Decision Trees and Prosodic Modification Costs," Proceedings of *Eurospeech*, 1999.

[4] Wang R. H., et al., "Towards a Chinese Text-to-Speech System with Higher Naturalness," Proceedings of *ICSLP*, 1998.

[5] Yi, J. and J. Glass, "Natural Sounding Speech Synthesis using Variable-length units," Proceedings of *ICSLP*, 1998.

[6] Balestri M. et al., "Choosing the Best to Modify the Least: A New Generation Concatenative Synthesis System," Proceedings of *Eurospeech*, 1999.

[7] Meng, H. et al., "Preliminary Developments towards a trilingual speech interface for financial information inquiries," Proceedings of the *International Symposium of Signal Processing and Intelligent Systems*, 1999.

[8] Meng, H. et al., "CU FOREX: A Bilingual Hotline for Foreign Exchange Enquiries," Proceedings of the *International Symposium of Signal Processing and Intelligent Systems*, 1999.

[9] Taylor, P. et al., "The architecture of the Festival speech synthesis system", in *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pp.147-151.

[10] Linguistic Society of Hong Kong (香港語言學會), *Hong Kong Jyut Ping Character Table (粵語拼音字表)*, Linguistic Society of Hong Kong Press, 1997.

[11] Lee, T. et al., "Microprosodic control for Cantonese Text-to-Speech Synthesis," Proceedings of *Eurospeech*, 1999.