Synthesis Units for Conversational Speech — Using Phrasal Segments (PartII) *

◎ Nick Campbell, ATR

1 Introduction

This paper is the second in an open-ended series that discusses the needs and possibilities of conversational speech synthesis. It is based on our experience with a very large corpus of spontaneous conversational speech, collected as part of the JST/CREST Expressive Speech Processing Project [1].

The corpus was produced with the aid of volunteer subjects who wore head-mounted studio-quality microphones throughout their ordinary working life to record their everyday spoken interactions over a period of five years.

A previous paper [2] introduced the prototype Chakai conversational-speech-synthesis interface and detailed a method for selecting phrase-sized speech segments from a conversational-speech corpus according to a constraint-based framework which incorporated Self, Other, and Event (i.e., speech- and/or discourse-act) parameters.

The present paper discusses problems arising from the use of that interface with the intention of clarifying the needs of conversational speech synthesis for use in human-to-human (e.g., speech translation), robot-to-human, or information-service (e.g., customer-care) interactions.

It shows that the text-to-speech paradigm prevalent in current speech synthesis technology has serious failings, and suggests that (as with speech recognition in the past) the issue can be resolved by use of an improved language-model, in this case requiring a mapping from text-based language to the types of spoken-language that are more appropriate to conversational interactions.

2 Chakai-Plus

We have previously distinguished two types of utterances found in conversational speech; differentiating the predominently I-type (which convey propositional content or ‘information’) from the predominently A-type (which serve primarily to convey speaker emotions, discourse-intentions and ‘affect’) [3]. Accordingly, the previously presented (A-type) Chakai conversational-speech synthesis interface has been amalgamated with a conventional chatr-type speech synthesiser, using the same voice of the corpus speaker, to enable free input from text as well as whole-phrase unit-selection based on discourse constraints.

There is a noticeable difference in quality when whole-phrase utterances are integrated with utterances produced by phone-level unit-selection in a conversation, but the functional effect is adequate. The chatr-type synthesis is sufficient for the I-type utterances but whole-phrase synthesis is necessary for the A-type discourse units which require finer and more subtle expression of prosody and voice quality.

In order to improve the overall acceptability of this hybrid system, we would like to be able to select more phrase-sized segments from the corpus for use intact in I-type utterances. In testing the combined interface, we have observed that many of the I-type utterances currently generated by phone-level unit concatenation already exist in similar form in the corpus. However, they are not detected in a text-based search because they are not an exact match with the input text sequence.

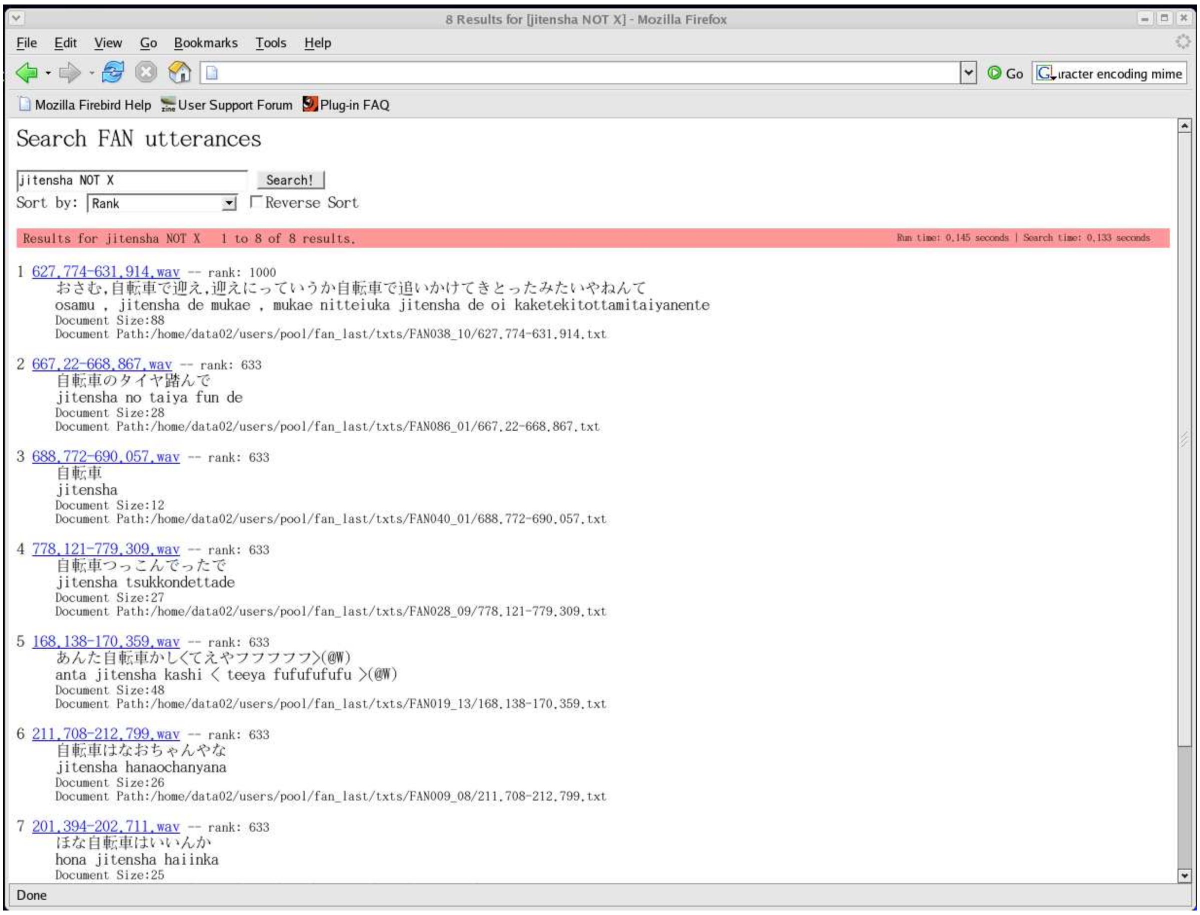
This difference can be illustrated (in English) as follows: the user types “I want to go to Kyoto” and the system carries out a text-based search for tokens of the same utterance in the corpus. Nothing is returned so the text is passed to the chatr-module for phone-level synthesis. In fact there may be several examples of the utterance “I *wanna* go to Kyoto” (or its Japanese equivalent) in the corpus, since this is how it is normally produced in conversational speech, but that exact text sequence was not entered by the user (who is normally not consciously aware of these speech mannerisms) so no matching tokens were found and the synthesised result is less natural in two respects; both acoustically (as a result of the concatenation) and also in respect to conversational appropriacy.

3 Incorporating Swish-E

On the assumption that given five years of conversational speech utterances, a large portion of those can be re-used, partially or intact, as units for synthesis in a future conversational speech system, we need a way of searching the corpus for the most appropriate candidate utterances which have a meaning close to the intended utterance. These can then be used whole or in part as synthesis units to reduce phone-level concatenation.

In an attempt to resolve this data-retrieval problem, the public-domain, open-source, web-browser search-engine software SWISH-E [4] has been incorporated as a front-end for selecting utterances from the speech corpus. This search engine provides a Google-like index of all utterances in the corpus and

* 対話音声合成、 フレーズ単位の利用について
ニック キャンベル ATR Media Information Science,
Keihanna Science City, Kyoto, 619-0288, Japan,


8 Results for [jitensha NOT X] - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
ıracter encoding mime
Mozilla Firebird Help
User Support Forum
Plug-in FAQ
Search!
Done

Search FAN utterances

jitensha NOT X

Sort by: Rank □Reverse Sort

```
Results for jitensha NOT X   1 to 8 of 8 results.          Run time: 0.145 seconds | Search time: 0.133 seconds

1 627.774-631.914.wav -- rank: 1000
    おさむ,自転車で迎え,迎えにっていうか自転車で追いかけてきとったみたいやねんて
    osamu , jitensha de mukae , mukae nitteiuka jitensha de oi kaketekitottamitaiyanente
    Document Size:88
    Document Path:/home/data02/users/pool/fan_last/txts/FAN038_10/627.774-631.914.txt

2 667.22-668.867.wav -- rank: 633
    自転車のタイヤ踏んで
    jitensha no taiya fun de
    Document Size:28
    Document Path:/home/data02/users/pool/fan_last/txts/FAN086_01/667.22-668.867.txt

3 688.772-690.057.wav -- rank: 633
    自転車
    jitensha
    Document Size:12
    Document Path:/home/data02/users/pool/fan_last/txts/FAN040_01/688.772-690.057.txt

4 778.121-779.309.wav -- rank: 633
    自転車つっこんでったで
    jitensha tsukkondettade
    Document Size:27
    Document Path:/home/data02/users/pool/fan_last/txts/FAN028_09/778.121-779.309.txt

5 168.138-170.359.wav -- rank: 633
    あんた自転車かし<てえやフフフフフ>(@W)
    anta jitensha kashi < teeya fufufufufu >(@W)
    Document Size:48
    Document Path:/home/data02/users/pool/fan_last/txts/FAN019_13/168.138-170.359.txt

6 211.708-212.799.wav -- rank: 633
    自転車はなおちゃんやな
    jitensha hanaochanyana
    Document Size:26
    Document Path:/home/data02/users/pool/fan_last/txts/FAN009_08/211.708-212.799.txt

7 201.394-202.711.wav -- rank: 633
    ほな自転車はいいんか
    hona jitensha haiinka
    Document Size:25
```

Fig. 1 The SWISH-E interface to the JST/CREST Expressive Speech Corpus

enables refined retrieval of selected utterance samples by the use of AND and NOT conditions in the search key (figure 1). The result is an extremely fast, currently web-based, 'speech synthesis' interface.

For example, searching a corpus of 436,961 utterances from one speaker, using "mama AND papa NOT X" (i.e., look for all sentences containing both the word 'mama' and the word 'papa' but exclude all those that contain the symbol 'X' (which indicates a noisy recording)) as the search-key, yielded 4 results, with a run time of 0.077 seconds and a search time of 0.065 seconds on a notebook PC. A search for 'honma NOT X" took 0.094 seconds (search-time 0.083 secs) and yielded 2498 results which were presented 15 to a page in Google-style layout, showing the Japanese text, its Kakasi-produced romaji equivalent, and the full path to the speech file with a click-to-listen link for each utterance displayed.

Utterances can be further filtered, by an intermediate programme, according to their acoustic characteristics in order to constrain the search for an extended Chakai-style A-type interface.

Given the improved access facilities described above, the remaining problem is to produce an efficient language-model so that a closer approximation to the idiomatic and colloquial nature of the corpus contents can be produced automatically. This largely remains as future work.

4 Discussion

This paper does not maintain that the collection of a five-year corpus is necessary in order to produce a conversational speech synthesiser, but rather that the effort is essential (and perhaps needs only to be be performed once) if we are to understand, in an objective way, the needs of such conversational-speech synthesis.

It has become clear from our analysis of this corpus that the types of utterance found in normal everyday conversational speech differ greatly from those found in text. It will be necessary to produce a grammar or language-model, which maps from the ideal context-neutral sentences that we have become accustomed to writing, in order to produce a text sequence closer to that of the colloquial idiom so as to reduce the search distances when browsing the corpus or searching for utterance tokens.

References

[1] JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.atr.jp/esp

[2] Synthesis Units for Conversational Speech — Using Phrasal Segments, Proc ASJ Autumn Meeting 2004.

[3] Getting to the Heart of the Matter; Speech as the Expression of Affect, **Language Resources and Evaluation**, Volume 39, Issue 1, pp. 111-120, 2005

[4] SWISH-E — Simple Web Indexing System for Humans, Enhanced Version: http://swish-e.org/