Selecting Speech Fragments for Affect Display in Concatenative Expressive Speech Synthesis *

◎ Nick Campbell, NiCT/ATR

1 Introduction

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

For this paper we examined aspects of the dialogue structure of Japanese speakers in telephone conversations with male, female and family interlocutors. The conversations were recorded over a period of several months and each lasted approximately 30 minutes. The corpus is part of subset ESP-C of the JST/ATR Expressive Speech Corpus.

Figure 1 lists the hundred most common "words" found in the corpus. Very few would be translated in a conventional speech translation system, being considered more as 'noise' than 'signal'. Figure 2 shows a section of one conversation, plotting the speech/non-speech activity across time for each partner. It is clear from the figure that there is considerable overlap and much "fragmentation" of the speech, with turns progressively alternating but not in the strict on/off manner expected by many dialogue system interfaces.

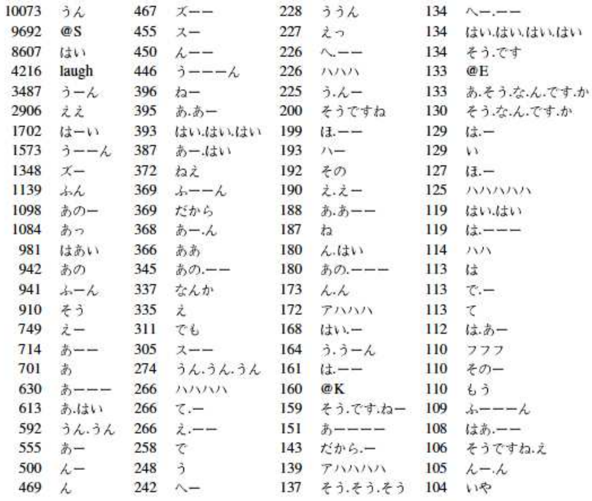
It has been argued elsewhere [2] that this fragmentation, caused by the frequent insertion of affective grunts, is used to indicate speaker-listener relationships throughout the discourse, to signal discourse-control information, and to show speaker state(s).

In this paper, we claim that these repetitive and very frequent fragments can be used to increase naturalness in the synthesis of expressive conversational speech such as might be required of a domestic robot, a customer-friendly information service, toys, or natural conversational speech translation.

2 Fragmentation & Dialogue Flow

Figure 2 is part of a screen dump of the web-based interface to the corpus, where by mousing over the sections, the text of the speech they represent can be interactively displayed with the audio. For a numerical analysis of the frequency of these affect-grunts in the conversational dialogues, we prepared a computer program to distinguish "linguistic content" from "non-verbal speech" in the transcriptions.

Clearly this distinction is not unambiguous, and ideally some human intervention would be required to distinguish e.g., "ano" used as a determiner from "ano" used as a hesitation marker, in the absence of punctuation. However, we used the dictionary shown in figure 1 in conjunction with Mecab's part-of-speech analysis [3] to detect fillers and interjections, and separated the text into affective (A-type) and linguistic content (I-type) components.



10073	うん	467	ズーー	228	ううん	134	へー.ーー
9692	@S	455	スー	227	えっ	134	はい.はい.はい.はい
8607	はい	450	んーー	226	へ.ーー	134	そう.です
4216	laugh	446	うーーーん	226	ハハハ	133	@E
3487	うーん	396	ねー	225	う.んー	133	あ.そう.な.ん.です.か
2906	ええ	395	あ.あー	200	そうですね	130	そう.な.ん.です.か
1702	はーい	393	はい.はい.はい	199	ほ.ーー	129	は.ー
1573	うーーん	387	あー.はい	193	ハー	129	い
1348	ズー	372	ねえ	192	その	127	ほ.ー
1139	ふん	369	ふーーん	190	え.えー	125	ハハハハハ
1098	あのー	369	だから	188	あ.あーー	119	はい.はい
1084	あっ	368	あー.ん	187	ね	119	は.ーーー
981	はあい	366	ああ	180	ん.はい	114	ハハ
942	あの	345	あの.ーー	180	あの.ーーー	113	は
941	ふーん	337	なんか	173	ん.ん	113	で.ー
910	そう	335	え	172	アハハハ	113	て
749	えー	311	でも	168	はい.ー	112	は.あー
714	あーー	305	スーー	164	う.うーん	110	フフフ
701	あ	274	うん.うん.うん	161	は.ーー	110	そのー
630	あーーー	266	ハハハハ	160	@K	110	もう
613	あ.はい	266	て.ー	159	そう.です.ねー	109	ふーーーん
592	うん.うん	266	え.ーー	151	あーーーー	108	はあ.ーー
555	あー	258	で	143	だから.ー	106	そうですね.え
500	んー	248	う	139	アハハハハ	105	んー.ん
469	ん	242	へー	137	そう.そう.そう	104	いや

Fig. 1 Counts of the hundred most common utterances of Japanese, as found in the ESP corpus of natural conversations. All function to display affect

Table 1 shows a count of I-type and A-type utterances for a male and a female speaker according to type of partner. The male (JMC) data included 9,056 novel words, 58,754 close repeats, 52,399 far repeats in a total of 20,459 utterances (where an 'utterance' is defined as a stretch of speech not including a pause longer than 300 msec). Here, 'novel words' are nouns, verbs, or proper names that serve a strictly lexical function; i.e., the words that would persist in a clipped telegraph-type rendering of the utterance. 'Close repeats' are words (or morphemes) that are repeated more than a threshold number of times within a distance of 100 morphemes of each other. 'Far repeats' are those that are also repeated but with a minimum distance of 100 morpheme units between each repeat. The female speaker's data (JFC; 26,119 utterances) included 11,482 novel words, 71,498 close repeats, and 61,258 far repeats.

Some of these repeats will be syntactically determined, but many, especially the close repeats, are characteristic of conversational speech styles. We can see from the organisation of speech fragments in Figure 2 that turn-taking in conversational speech is not like a game of tennis, where there is only one

Table 1 Counts of utterance type per partner

speaker	female	(JFC)	male	(JMC)
content	I-type	A-type	I-type	A-type
to female	30,079	31,897	17,227	25,605
to male	33,068	35,197	28,483	27,264
to family	20,069	33,246	25,372	27,518

* 波形接続方式の対話音声合成に起ける発話歎異選択について
ニック キャンベル NiCT/ATR-SLC, Keihanna Science City, Kyoto, 619-0288, Japan,

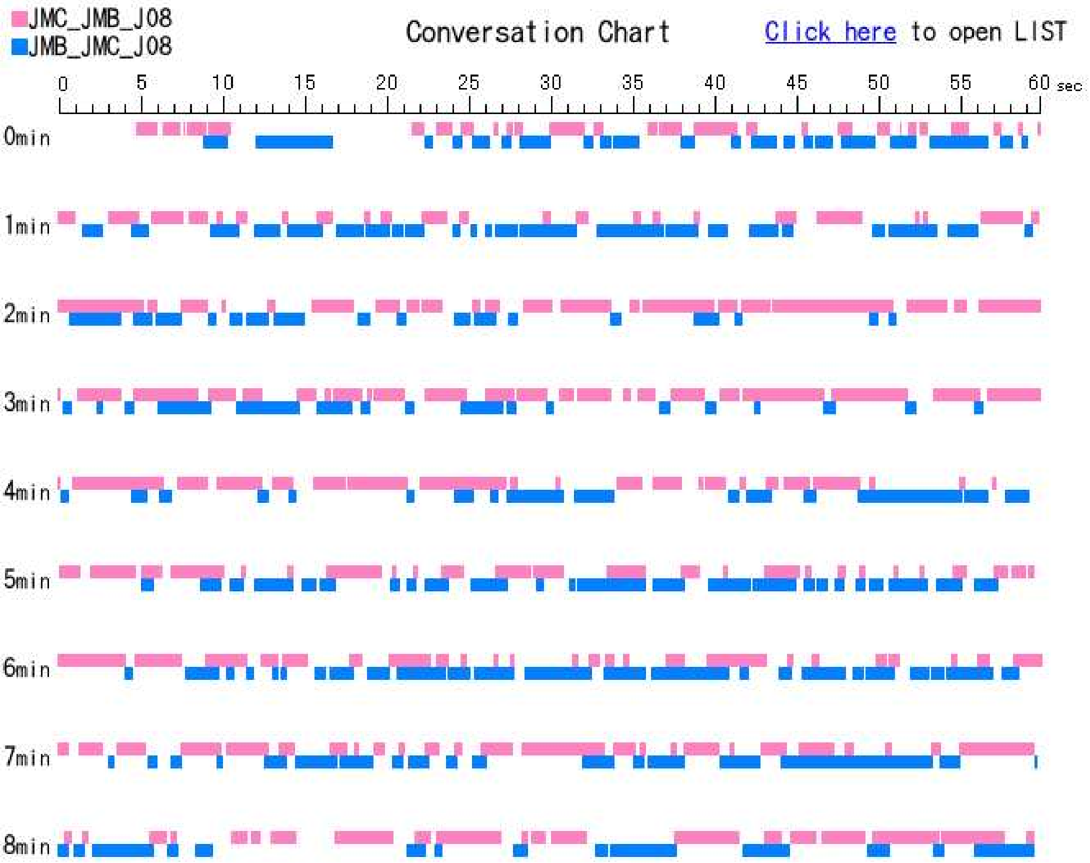


Fig. 2 Speech & silence plots for the first 9 minutes of conversation 8 between two male speakers, JMC and JMB, showing fragmentation of the discourse and progressive but not absolute alternations of speaker dominance. Each line shows one minute of speech, with speaker JMC's speech activity plotted above and that of speaker JMB plotted below. White space indicates lack of speech activity

ball that is passed from one partner to the other. Although it is usually clear in most parts of the figure who is the 'dominant' speaker at any given moment, there is considerable overlap, and much simultaneous speaking. Yet when listening to this dialogue, the impression is one of harmony rather than discord. The overlaps are boosting, affirmative, encouraging, and supportive.

3 Selection of Units

In conversational speech synthesis, it will be necessary to emulate this behaviour, which is a form of 'active listening', and to synthesise the laughs, verbal nods, and other affective displays, in order to provide the expected support for the speaker.

Because almost all of these affective grunts are pause-delimited in the speech, there is no need to consider a join cost in the unit selection. However, because they indicate sensitive interpersonal relationships rather than syntactic phrasing or semantic relations, the target cost becomes increasingly important. This problem can be overcome by the use of acoustic features in their selection.

We have shown that a principal-component reduction of a set of 14 acoustic measure correlates well with changes in affective state and relationship with the interlocutor. By using this measure as an indication of intensity of affect, it is possible to select fragments to fit the desired tone of the conversation.

4 Conclusion

It has been shown that active listening results in considerable speech overlaps in natural conversations and it is claimed that simulation of these utterances will increase the naturalness of conversational speech synthesis. Certain discourse fragments are frequently repeated yet vary considerably in prosody and phonation style. They can be selected by consideration of these acoustic characteristics. Currently, we are producing a language model by which we can select between them to insert appropriatel fragments into the stream of speech.

References

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

[2] 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

[3] MeCab: Yet Another Part-of-Speech and Morphological Analyzer — http://mecab.sourceforge.jp/

[4] Conversational Speech Synthesis and the Need for some Laughter, **IEEE Transactions on Audio, Speech, and Language Processing** V.14, N.4, 1171-1178, 2006.