Expressing Affect in Spontaneous Japanese Conversational Speech [*]

◎ Nick Campbell, ATR

1 Introduction

This paper presents data to support the contention that the supposedly "ill-formed" structure of spontaneous Japanese speech actually provides a mechanism whereby the speaker can express both propositional content *and* affective information simultaneously in the same utterance.

Specifically, it introduces a notion of "wrappers" and "fillers" wherein 'content-rich' sections of speech are interspersed with 'affect-rich' discourse and interpersonal markers. The wrappers typically found at the start and end of each utterance provide frequent and standardised reference points by which a listener can make an affective judgement about the states and intentions of the speaker and the progress of the discourse.

This work is based upon an analysis of a very large corpus of *really* spontaneous Japanese conversational speech, collected as part of the JST/CREST Expressive Speech Processing Project carried out at ATR between the years of 1999 and 2005 [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 the five-year period. All utterances were transcribed by hand at the phonemic level. Because the volunteers wore the recorders over a very long period of time, and in a variety of interpersonal situations, the speech is highly natural and *very* spontaneous.

Table 1 The most common words in the corpus, (one speaker, numbers show occurrence frequency)

48038	うん	1733	で	829	ま
15555	あ	1675	ほんで	800	んんん
10961	ふん	1550	うんうん	787	まあ
8408	うーん	1535	もう	751	わかった
7769	え	1428	でも	737	や
5796	ああ	1422	ふんで	730	ありがとう
4891	ほんま	1412	はあ	713	あれ
4610	あー	1370	ええ	703	そうそうそう
3704	んん	1329	そう	692	は
3608	はい	1299	ふんん	692	そうなんや
3374	なんか	1291	ほんまあ	687	あたし
3164	ん	1246	うんうんうん	679	んんーん
3010	いや	1227	あのう	674	はいはい
2942	ふーん	1206	ううん	673	そうそうそうそう
2860	あの	1118	これ	658	フフ
2246	ふうん	1108	そうそう	645	せやな
2238	なあ	1085	おん	623	ほんなら
1871	そうなん	1079	まあな	599	うんうんうんうん
1761	な	903	あああ	588	ほん
1736	うんん	871	だから	583	よいしょ

2 Wrappers & Fillers

For speech synthesis purposes [2, 3], we have previously distinguished two types of utterance found in conversational speech, i.e., differentiating the predominently *I-type* (which serve to convey propositional content or *'information'*) from the predominently *A-type* (which serve primarily to convey speaker emotions, discourse-intentions and to express *'affect'*) [4]. See table 1 for examples of the latter with counts of their occurrence for one speaker in the corpus.

However, to contend that any single utterance must function primarily as either A-type or I-type is clearly an oversimplification, as both types of information are often signalled simultaneously. This paper extends the distinction to explain how a mixture of the two types of information creates the so-called "ill-formedness" that is considered characteristic of spontaneous interactive speech.

Whereas in written communication the word sequences are usually carefully deliberated, in spontaneous speech, the flow is generated in real-time and a stream of words and phrases would typically (in colloquial English) appear as follows:

" ... erm, anyway, you know what I mean, ..., it's like, er, sort of **a stream of** *... er ...* **words, and phrases**, *all* **strung together**, *if you know what I mean, you know ... "*

where the words in bold-font form the content (the filling of the utterance) and the italicised words form the wrapping or decoration around the content.

Here the term 'filler' is used to describe the I-type content (that which would normally be included in an orthographic transcription), and the term 'wrapper' is used to describe the A-type portions of the utterance, that are often considered as ill-formed.

This usage is in (deliberate) contrast to the usual interpretation of a 'filler' as something which occupies a 'gap' or empty space in an utterance. On the contrary, this paper suggests that by their very frequency, these non-propositional, often non-verbal, speech sounds provide not just time for processing the spoken utterance but also a regular base for the comparison of fluctuations in voice-quality and speaking-style.

3 Longest Common Substrings

The definition of a 'word' in any language is very difficult, but we need to define a suitable unit for phrase-level concatenative synthesis from a very large corpus. These frequent wrappers can be of

[*] 自然対話音声の情動情報表現について

ニック キャンベル Department of Cognitive Media Informatics,
ATR Media Information Science Labs, Keihanna Science City, Kyoto, 619-0288, Japan,

あ, もしもし，あのちょっと けいやくのないようへんこうしていただきいんですけど
#（もしもし。契約の内容を変更していただきたいのですが）
しらんゆう**ねんな**, ひつこい**ねん もう** ぴかぴかぴかぴかひかってるからきになってさあ
#（知らないと言っているのに、しつこいぴかぴか光っているので気になって）
たべれん**ねん**で, たべれん**ねんけど**きもちわるいし, まだまだしんどいし ,**みたいな**
#（多分食べられます。食べられるのですが、気持ちが悪いしまだしんどい、と言った感じで）
だかほんま あんまたたかんでいいらしい**ねんけど**, **ま，** ちょっとほこりおとすていど
#（だから本当に、あまり叩かなくて良いらしいです。少し埃を落とす程度で）
うんうんうん ,でもさ, どうせさ, いろいろあつめんねやったら , これをしってたら
#（どうせ色々集めるのなら、これを知っていれば）
うん, **そら，** こまま, こおりやまのほうちょっとまっすぐいったところ**やねん けどな**
#（はい。このまま郡山の方へまっすぐ言ったところなのですが）
まあ はんなどうろあるから **なあ** ,やっぱりつうこうりょうすくないかもしれん **よなあ**
#（まあ、阪奈道路があるからやっぱり交通量は少ないかもしれませんね）
それもかんがえようよな, **なんか** ほんまにきんてつでぜんぶすんねやったらいいけど
#（それも考えようですよ。本当に近鉄で全部するのならいいけれど）
あるくのいたい, **む, なんか** どっちかはんぶんがすごいしびれてあるかれへん**ねん**て
#（歩くのが痛い。どちらか半分がとても痺れて歩けないのです)
なんかめんどくさいな, おかしつねにかっとかなあかん**やん とか** おもっとってんけど
#（何か面倒くさいなあ。お菓子は常に買っておかなければならないと思っていたのですが）
なんかさあ，あの かたちがちゃんとなってへんからはきにくいすりっぱってある **やん**
#（形がちゃんとしていないために歩きにくいスリッパがあるじゃないですか。）

Fig. 1 Sample utterances having a length of between 20 and 40 mora, selected at random from the corpus. Each utterance is followed by its equivalent in standard Japanese for comparison. Bold font shows the 'wrappers' in these utterances

great use in locating the boundaries of the filler portions and for thereby segmenting the corpus into smaller and more useful phrase-sized units.

In order to produce a dictionary of frequent wrappers without resource to linguistic knowledge, we used a longest-common-substring algorithm to identify the most frequent kana sequences in the corpus. As training data, we used the set of transcribed utterances having a length of between 20 and 40 kana (n=43186). By sorting and matching characters from left-to-right, we obtained the utterance-initial forms (n=899), and matching right-to-left obtained the utterance-final forms (n=957). The sequences thus detected also appeared very frequently within these longer utterances and so provide candidate segmentation points (see Figure 1).

By chunking the longer uttterances in this way, we were able to greatly extend the usable phrases and utterance subsections for the concatenative speech synthesis described in [2] and [3].

4 Remaining Work

In order to make use of these 'wrappers' to improve the spontaneous nature of synthesised speech, we now need to produce an inventory of their functional equivalence. There are many types that are different in segmental composition but which function equivalently in the discourse (see Table 1). Collating these into usable classes remains as future work.

5 Discussion

This paper proposes that the evolution of this supposedly "broken" form of speech is not just a side-effect of poor performance in real-time speech generation processes, but that the inclusion of frequently repeated non-content segments allows the speaker to use them as carriers for affective information such as is signalled by differences in voice quality and speech prosody. Their high frequency (and relative transparency with respect to the propositional content) allows small changes or contrasts in phonation style to be readily perceived by the listener. For efficient detection of these discourse effects by machine, we should perhaps focus on the highly-frequent, so-called disfluent sections of speech alone.

Acknowledgements

This work is partly supported by the Japan Science & Technology Corporation (JST), partly by the National Institute of Information and Communications Technology (NiCT). The author is grateful to the management of ATR, especially to the Media Information Science Laboratory (MIS) for continuing encouragement and support.

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] Synthesis Units for Conversational Speech — Using Phrasal Segments - Part II, Proc ASJ Autumn Meeting 2005.

[4] 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