On the Structure of Spoken Language

Nick Campbell

Advanced Telecommunications Research Institute
Media Information Science Laboratory, Department of Cognitive Media Informatics,
Keihanna Science City, Kyoto, Japan 619-0288
nick@atr.jp

Abstract

The special structure of spoken language is often described as "ill-formed" but this paper shows that it is ideally suited to the simultaneous expression of (a) propositional content (*i.e., linguistic information*) and (b) speaker-state, discourse management cues, and speaker-listener-relationships (*i.e., affective information*). This paper shows that by the frequent insertion of so-called "fillers" and other repetitive fragments, the speaker provides the listener with constant reference points for evaluating affective states as displayed by voice-quality information.
Keywords expressing affect, spontaneous communication, 'wrappers & fillers', sentence structure, discourse control

1. Introduction

Previous work [1] has shown that non-lexical fragments are extremely common in conversational speech. From analysis of 150,000 transcribed conversational utterances, recorded from one speaker over a period of four years, we found almost 50% to be non-lexical; i.e., they could not be adequately understood from a transcription of their text alone. (Table 1 provides detailed figures, Table 3 shows some examples). Very few of these utterance types can be found as an entry in a standard language dictionary, yet it was confirmed that the intended meanings of many of these non-verbal utterances (or conversational 'grunts') can be perceived consistently by listeners even when presented in isolation without any discourse context information. In many cases, the intentions underlying the utterances can be appropriately and consistently paraphrased even by listeners of completely different cultural and linguistic backgrounds [2]. This paper extends the analysis to include disfluent fragments in longer utterances, and offers an explanation for the so-called 'ill-formed' nature of spontaneous speech.

We have previously shown that the voice quality (i.e., mode of laryngeal phonation) of these utterances varied consistently and in much the same way as (but independently of) fundamental frequency, to signal paralinguistic and affect-related information [3]. The mode of laryngeal phonation can be measured from an estimate of the glottal speech waveform derivative (a result of inverse filtering of the speech using time-varying optimised formants to remove vocal tract influences [4]) by calculating the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle-to-cycle minimum derivative [5]. In its raw form it is weakly correlated with the fundamental period of the speech waveform ($r = -0.406$), but this can be greatly reduced by $NAQ = log(AQ) + log(F_0)$, to provide an uncorrelated ($r = 0.182$) Normalised Amplitude Quotient (henceforth 'NAQ') [6].

We showed that the factors 'interlocutor', 'politeness', and 'speech-act' all had significant interactions with this variation [7]. The factor 'interlocutor' was analysed for NAQ and F_0, grouped into the following classes: Child (n=139), Family (n=3623), Friends (n=9044) Others (n=632), and Self (n=116). It is clear from figures 1 and 2 that both F_0 and breathiness are being controlled independently for each class of interlocutor. Repeated t-tests confirm all but the child-directed (n=139) voice-quality differences to be significant at $p < 0.001$. Figure 1 shows median NAQ and F_0 for the five categories of interlocutor. NAQ is highest (i.e, the voice is breathiest) when addressing strangers (politely), and when talking to children (softly). Self-directed speech shows the lowest values for NAQ, and speech with family members exhibits a higher degree of breathiness (i.e., it is softer) than speech with friends. F_0 is highest for child-directed speech, and lowest for speech with family members (excluding children). Figure 2 shows results for family-directed speech in more detail, and shows that family members can be ordered by voice-quality settings as follows: $daughter > father > nephew > mother = older\ sister > aunt > husband$. This reflects the view that increased breathiness indicates a higher degree of 'care' taken in the speech.

Table 1: Counts of non-verbal utterances in the transcriptions for one speaker in the ESP corpus. Utterances labelled 'non-lexical' consist mainly of sound sequences and combinations not found in the dictionary, but may also include common words such as "yeah" "oh", "uhuh", etc.

total number of utterances transcribed	148772
number of unique 'lexical' utterances	75242
number of 'non-lexical' utterances	73480
number of 'non-lexical' utterance types	4492
proportion of 'non-lexical' utterances	49.4%

2. A 3-Dimensional Framework

To account for the above effects, we proposed a 3-dimensional framework for the categorisation of a speech utterance that also serves for the detailed specification of an utterance in speech synthesis [8]. It was assumed that speakers and listeners must share a protocol for the communication of affective information that can be interpreted in place of, or in conjunction with, the more well-formed semantic utterances that are produced for the communication of propositional content. i.e., that the listener can interpret a conversational grunt *in ways that the speaker apparently intended* implies that the current assumption of spoken communication as consisting largely of semantic elements func-

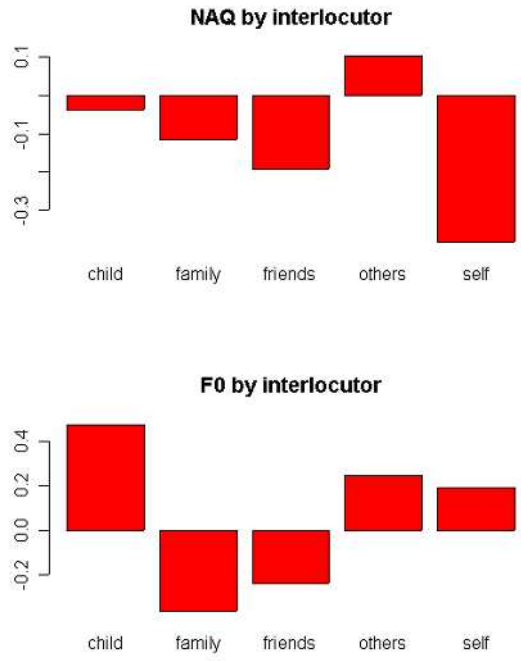

Figure 1: *Median values of NAQ and F_0 plotted for interlocutor. The data are (z-score) scaled, so values are in SD units. 0 represents the mean of the distribution*

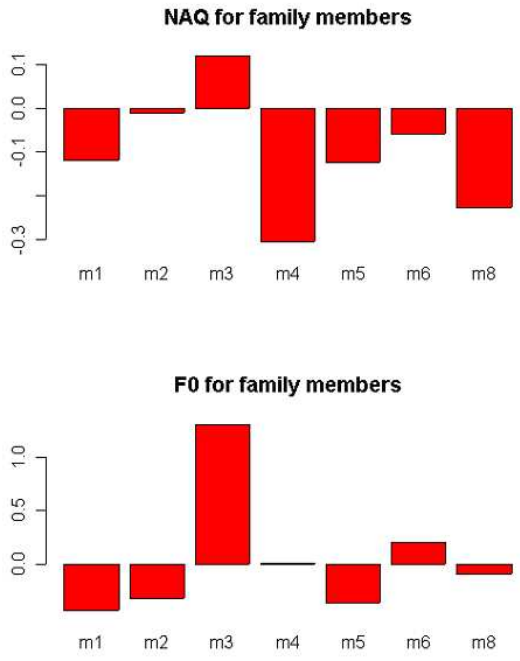

Figure 2: *Median values of NAQ and F_0 for family members. m1: mother, m2: father, m3: daughter, m4: husband, m5: older sister, m6: sister's son, m8: aunt*

tioning within a syntactic framework (as rendered in a cleaned-up linguistic transcription of the speech content) is inadequate for describing the full function of spoken language.

To better account for the speaking-style and phonation characteristics of an utterance, we need to know not just what is said, but also who is talking to whom, where, and why. This information can be coded in higher-level terms as a combination of the following three features or 'SOE' constraints: (i) Self, (ii) Other, (iii) Event, as in (1), which defines an utterance (U) as specified by the pair *self* (S) and *other* (O) given *event* (E):

$$U = (S, O)|E \tag{1}$$

where the feature *Self* can take different values (representing *strong* and *weak* settings with respect to the dimensions *mood* and *interest* respectively) and the feature *Other* can also take different values (representing *strong* and *weak* settings with respect to the dimensions *friend* and *friendly* respectively), and the feature *event* represents a speech act (in a wider and more detailed sense than Searle defined) or a discourse move.

The feature *Self* refers to (a) the state of the speaker and (b) his or her interest in the content of the utterance. For example, a healthy, happy, person is likely to speak more actively than an unhealthy or miserable one. One who is interested in the topic or highly motivated by the discourse is likely to be more expressive than otherwise.

The feature *Other* refers to (a) the relationships between speaker and hearer, and (b) the constraints imposed by the discourse context. A speaker talking with a friend is likely to be more relaxed than when talking with a stranger, but will also probably be more relaxed when talking informally, e.g., in a pub, than when talking formally, e.g., in a lecture hall.

The feature *Event* minimally requires that we distinguished each utterance as being either of I-type or A-type content; the former primarily expressing propositional content (or *Information*, and the latter primarily expressing *Affect*. Since the transfer of such information can be bi-directional, we also distinguish *giving* from *getting* (see figure 3).

For simplicity, the figure shows each dimension as having four 'settings' (which has proved useful for speech synthesis [8]), and reduces discourse intentions to a two-by-two matrix, but clearly more detailed explanation is still required for a full account of speaking-style controls,. In our tagging of the conversational speech corpus, each utterance is first categorised in terms of its directionality, then in terms of modality, i.e., whether primarily of I-type or of A-type, as in Table 2, and then for affective subcategory if relevant. All A-type utterances, whether lexical or grunts are candidates for an affect label.

Table 2: Basic utterance types for the *Event* category

	seeking	offering
I-type	interrogative	declarative
A-type	back-channel	exclamative

3. Wrappers & Fillers

To contend that any single utterance must function primarily as either A-type or I-type is clearly an oversimplification, since both types of information are often signalled simultaneously. This section extends the above 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.

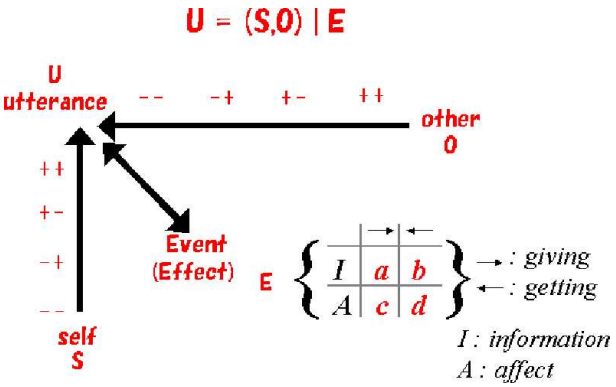

Figure 3: *A 3-dimensional framework for categorising speaking-style and expressiveness; an utterance is realised within the constraints of speaker (self) and interlocutor (other) according to the discourse intention (event). Voice quality, speech-rate, prosodic range, and voice power will vary accordingly*

Table 3: The most common complete utterances in the corpus, (data from one speaker, numbers show occurrence frequency). Note the highly repetitive nature of these common expressions

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

Whereas in written communication the word sequences are usually carefully deliberated and well-formed, in the case of spontaneous speech the flow is generated in real-time and a stream of words and phrases might 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 (or 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 (the text which would normally be included in a cleaned-up orthographic transcription of the utterance), and the term 'wrapper' is used here 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 supposed empty space in a discourse. On the contrary, this paper suggests that by their very frequency, these non-propositional and 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.

4. Chunking I-type Utterances

The biggest difference that is immediately apparent between I-type and A-type utterances is their length. Although some longer utterances such as "Good morning", "How are you today?", and "Did you see the game last night?!" can be considered as primarily phatic, and hence A-type, the transfer of propositional information that defines I-type utterances usually requires more words to be strung together in a longer sequence. For this paper, we examined the frequency of non-verbal fragments (including A-type affect markers) in these longer utterances, and segmented them further by thus distinguishing wrappers from fillers in the speech.

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 symbol sequences occurring at utterance-initial or utterance final positions in the transcribed corpus. As training data, we used the set of transcribed utterances having a length of between 20 and 40 kana characters ($n = 43,186$). A kana symbol in the Japanese phonetic alphabet approximately corresponds to a syllable. By setting a threshold of 10 repetitions as a minimum criterion for inclusion, and then sorting the utterances and matching characters from left-to-right to obtain the longest common substring, we obtained 899 frequently occurring utterance-initial forms, and then by matching right-to-left (i.e., by sorting the reversed strings) we obtained 957 frequent utterance-final forms.

These "edge-pattern" wrapper sequences were then matched wherever they occurred utterance-internally and were used as further segmentation points to divide the longer utterances into 'wrapper' and 'filler' sections, with the edge patterns being taken as wrappers and the intervening sections assumed to be 'fillers'. Figure 4 illustrates the result of this two-stage process. The 'words' in bold font being the common (typically non-lexical) 'wrappers'. Even to those who cannot read Japanese, it will be apparent from the figure that these are very frequent. Note that lines starting with a "#" are manually-produced transliterations and rarely include such terms. In a hand-checked subset of 1000 utterances we counted 2337 wrappers; an average of 2.34 per utterance. Note that single-character (single-syllable) wrappers are difficult to detect automatically without recourse to a morphological analysis of the transcription, so the actual number of occurrences may be much higher.

5. Discussion

Whereas the original purpose of this finer segmentation of longer utterances was to provide shorter units for use in a phrase-level waveform-concatenation speech synthesis system, we were struck by the frequency of A-type segments in what we assumed would be primarily I-type utterances. Our transcribers had been offered a 'yen-per-line' incentive to cut the utterances as finely as possible, but perceived these numerous longer ones as being single intonation units or difficult to segment more finely.

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。)

Figure 4: Sample utterances of Japanese conversational speech, selected at random from those having a length of between 20 and 40 mora in the corpus. Each utterance is followed by its equivalent transliteration in standard Japanese for comparison. Bold font shows the automatically-detected 'wrappers' in these utterances

If we return to the English example above, we see that "it's like", "er", and "sort of" form a sequence that might be perceived as a single prosodic unit, while actually only functioning minimally in a linguistic sense. Such words and phrases slip easily into idiomatic conversational speech, and allow both the speaker and the listener to reduce the cognitive processing load, but we believe that they also serve a more important function as highly-recognisable, frequently occurring segments by which the speaker may express affective information through use of subtle variations in voice-quality and other prosodic controls, and that the listener can use to judge speaker-state(s), speaker-listener relationship(s), and discourse control signals.

We therefore suggest that the evolution of this supposedly "broken" form of spontaneous 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, even if he or she is unfamilliar with the speaker.

6. Conclusion

This paper has presented a notion of "wrappers" and "fillers" wherein 'content-rich' sections of speech are interspersed with 'affect-rich' discourse and interpersonal markers. The A-type wrappers typically found at the start and end of each I-type content portion 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 supports the contention that the supposedly "ill-formed" structure of spontaneous speech actually provides a mechanism whereby the speaker can express both propositional content *and* affective information simultaneously in the same utterance.

7. Acknowledgements

This work is supported by the Japan Science & Technology Corporation (JST), the National Institute of Information and Communications Technology (NiCT), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan (SCOPE). The author is also grateful to the management of ATR for their continuing encouragement and support.

8. References

[1] Campbell, N., 2004. "Extra-Semantic Protocols; Input Requirements for the Synthesis of Dialogue Speech" in *Affective Dialogue Systems, Lecture Notes in Artificial Intelligence*, vol. 3068 Eds Andre, E.; Dybkjaer, L.; Minker, W.; Heisterkamp, P., New York, Springer. 221-228

[2] Campbell, N., & Erickson, D., 2004. "What do people hear? A study of the perception of non-verbal affective information in conversational speech", in *Jnl Phonetic Society of Japan*.

[3] Campbell, N., and Mokhtari, P., 2003. "Voice Quality; the 4th prosodic parameter", in Proc 15th ICPhS, Barcelona, Spain.

[4] Mokhtari, P., and Campbell, N., 2003. "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech" in *Trans IEICE Special Issue on Speech Information Processing*

[5] Alku P., and Vilkman, E., 1996. "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering", *Speech Comm*., vol.18, no.2, 131-138

[6] Alku, P., Backstrom, T., and Vilkman, E., 2002. "Normalized amplitude quotient for parametrization of the glottal flow", *J. Acoust. Soc. Am*., vol.112, no.2, 701-710

[7] Campbell, N., 2004. "Accounting for Voice Quality Variation", Proc 2nd Intl Conf on Speech Prosody, Nara, Japan.217-220

[8] Campbell, N., 2005. "Developments in Corppus-Based Speech Synthesis; Approaching Natural Conversational Speech" *IEICE Transactions on Information & Systems*, E88-D,3, 376-383

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