The Relationship between Sound and Meaning in Japanese Back-channel Grunts

Japanese title: Aizuchi no Onkyouteki Buhin to sorezore no Imi

Nigel Ward[1]
University of Tokyo

Abstract:

This version of the paper omits the Japanese abstract which appeared in the proceedings version (4th Annual Meeting of the (Japanese) Association for Natural Language Processing, March 1998, Fukuoka, pp 464–467.)

1 Motivation

Human vocalizations include not only words but also grunts. In conversation grunts play an important role in determining who will speak next and what sort of thing he will say. Interactions which take place without grunts are formal and rigid. For man-machine systems, this is acceptable today, when interactions are limited by the expectation that they will consist only of the reciprocal transmission of explicit factual questions and answers. Dialog systems which interact more flexibly or 'naturally' with humans, however, will have to be able to grunt and to understand grunts. This paper is a preliminary discussion of the sound patterns and meanings of grunts and some issues in their application.

2 Data

For a study of the timing of back-channel feedback, we recorded a corpus of a total of 80 minutes of Japanese conversation, consisting of (portions of) 18 conversations. Each conversation had two participants (Ward 1996; Ward & Tsukahara 1998).

This corpus contains a rich variety of grunts. In addition to the ubiquitous *un*, there is also *ee*, *hee*, *uu*, *uh*, *uun*, *ununun*, *huun*, *huh*, *hmmm*, *aa*, *na*, *oo*, *hoh*, *ooun*, *haan*, *eeen*, *hm*, *hhnh*, *ahaihaihai*, *sa*, *ya*, *hm-um*, and over a hundred more, with diverse prosody and voicing.

Most grunts appeared in one of four contexts, as seen in Table 1. Previous studies of grunts have tended to focus on full-turn grunts (Shinozaki & Abe 1997), and especially on the most expressive cases, interjections, where a single grunt may convey an entire sentence worth of disdain, annoyance, disgust, appreciation, joy, etc. Although salient, these are rare in normal conversation. In contrast, the current work has started with the most common type, back-channels, and been extended to also account for some of the other grunts.

Grunt Type	Example
Back-channel	A: *ikatta kara,* B: un, A: *betsu no onnanoko o* ...
Turn-opening or filler	A: ya *konshuu doyoobi nee* ...
Full-turn	A: *yarimakuten-no?* B: aa. A: *nan no tame ni?*
Post-completion grunt	A: *kirai dakara,* uun

Table 1: Some Places Where Grunts Occur

3 Hypotheses

Hypothesis A Grunts, unlike words, are not fixed sequences of phonemes, but are formed from various acoustic components.

Hypothesis B The various acoustic components of grunts individually bear meanings.

For example, nasalation indicates agreement, as in *un* versus *uu* and in *haan* versus *haa*; /m/ indicates contemplation, as in *um* versus *uu* and in *hm* versus *hh*; multiple syllables indicate a lack of anything to say, that is, a willingness to listen, as in *unun* versus *un*, and in *sososo* versus *so*. Other hypotheses appear in Table 2. These were arrived at by listening to the corpus, introspection, extrapolation from observations in the literature, and casual observation in daily life. None can be considered confirmed yet.

Hypothesis C The meaning of a combination of acoustic components is the combination of the

[1] nigel@sanpo.t.u-tokyo.ac.jp
Mech-Info Engineering, U. Tokyo, Bunkyo-ku Tokyo 113-8656
http:/www.sanpo.t.u-tokyo.ac.jp/∼nigel/
Thanks to Wataru Tsukahara for suggestions and to the Hayao Nakayama Foundation for support.

	Sound	Meaning	Examples etc.
a1	schwa	neutral	*uu*, *un*
a2	/a/	reticence, stalling	*aa*, *sa*, *ma*; minimal pair: *aa-so* vs. *so*
a3	/e/	agreement, sympathy, orientation to emotional content	*ee*, *ne*, *he*
a4	/o/	receipt of new information (Heritage 1984), orientation to factual content	*oo*, *hoh*, *ooun*
b1	nasalization	agreement	*un*, *haan*, *eeen*, *hhnh*, *huun*
b2	/m/	contemplation	*um*, *hm*, *maa*
b3	lip rounding	weightiness of information received	*ow*
b4	/j/	incipient desire to take a turn, desire to discourage the other from talking (Drummond & Hopper 1993)	turn-initial *ya*, *ja-nai-ya* of self-correction, topic closing *maa-ii-ya*
c1	breathiness and /h/	deference, politeness	*uh*, *huh*, *hee*, *hmm*, *ah*
c2	vocal fry	boredom, lack of involvement	
d1	pitch height	degree of interest (Shinozaki & Abe 1997)	
d2	rising pitch	incomplete understanding; invitations to continue	
d3	falling pitch	complete understanding; closure (Kawamori *et al.* 1995)	
e1	duration	amount of thought	
e2	number of syllables	lack of anything to add (Gardner 1997)	*unununununun*, *sososo*
e3	loudness	self-confidence, importance of utterance	
e4	abrupt end (sharp energy drop)	coldness, formality, haste	
f1	timing governed by speaker's cues	passivity	
f2	delayed timing	reticence, thinking	
f3	frequency of grunts	interest, attentiveness	

Table 2: Some hypothesized correspondences between the component sounds of Japanese grunts and their meanings. Spellings of examples are traditional rather than phonetically accurate, especially in the cases of *u*, *h* and *w*.

	Sound	Meaning	Examples etc.
a5	/i/	other-directness (vs. talking to oneself)	minimal pairs: *hai* vs. *ha*, turn-initial *iya* vs. *ya*, *dai* vs. *da*
b5	/s/	independent judgement (vs. passively believing the other)	minimal pairs: *saa* vs. *aa*, *soo* vs. *oo*
b6	/r/	immediate action required, or orientation to a extra-linguistic, real-world, happening	*ara*, *kora*, *are* of surprise
b7	/t/	willingness to control the conversation	minimal pairs: *eeto* vs. *ee*, *hontoo* vs. *hon*

Table 3: Hypothesized non-productive correspondences between the component sounds of Japanese near-grunts and their meanings.

meanings of each component (Kawamori *et al.* 1995; Takubo & Kinsui 1997).

This means that grunts are 'iconic', or, in other words, involve 'sound symbolism'. Sound symbolism in grunts appears to be a distinct from the onomatopoeic and mimetic systems of sound symbolism.

Hypothesis D The strength of an acoustic component in a grunt corresponds directly to the strength of the corresponding component of meaning.

That is, the components are not binary features, but graded. For example, degree of pitch upslope can distinguish between "I don't follow" completely and "I'm baffled". Also, syllabification (e2) varies from a sort of strong vibrato to complete separation into individual words. There may also be a continuum of possible pronunciations between the basic grunt vowels (a1–a4), with intermediate vowels having intermediate meanings.

Hypothesis E The order of acoustic components in grunts reflects the time course of the conversation and the time-course of thought.

For grunts which vary over time, it seems that sounds which appear early in the grunt tend to relate to the preceding utterance, whereas those which appear later relate to the upcoming dialog. For example, *na* seems to indicate strong agreement with the previous utterance, plus a neutral attitude as to whether the topic is worth continuing. Conversely, *an* seems to indicate mild skepticism with respect to the previous utterance, plus a willingness to listen agreeably to further utterances.

It also seems that changes in the sounds of the grunt reflect changes in the speaker's mental state. For example, the grunt *aun* seems to reflect the process of the respondent coming to understand and agree. Similarly, a grunt involving a rising then a falling pitch (forming a hat pattern) shows incomplete understanding, leading to, with more thought, complete understanding. This explains why this conjunction of pitch patterns indicates interest (Imaishi, cited in (Okada 1996)); if a person shows that he is actively thinking, then he is showing that he is interested.

4 Further Speculations

The distinction between grunts and words is not always clear-cut. Some 'words' often appear in the same contexts where grunts appear, and many of these seem to involve the same sound-meaning correspondences, as in *so*, *honto*, *naruhodo*, *uso*, *maji*, *ano*, *de*, *wa* and *yo*. Sentence-final particles, in general, seem to be grunt-like in these ways, as does laughter. There also seem to be some additional sound-meaning correspondences present in near-grunts (Table 3).

Moreover, hypotheses A though E also seem to apply to grunts in English, as do most of the sound-meaning correspondences of Tables 2 and 3. These can be seen, for example, in grunts and near-grunts such as *wow*, *yeah*, *nyeah*, *nright*, *alright*, *hm*, *uh-huh*, and *okay*. Of course, the frequencies of use of the components of grunts vary across languages, as do their pragmatic effects. For example, *yeah*, which indicates an incipient desire to take a turn, is often used as a back-channel in English, but *ya*, the analogous Japanese grunt, is not used in this way (White 1989); perhaps because it is considered impolite to show any desire whatsoever to take a turn while the other is still speaking.

5 Applications for Grunt Production

Conversations have many dimensions. One involves giving and receiving factual information and questions. Others include attitudinal, interpersonal, and emotional factors, where a participant indicates how pleased he is with the current topic, how interested he is, how much he likes the other participant, how much he approves of what he is saying, and so on (Shinozaki & Abe 1997). Yet another involves conversation control, where a participant indicates whether he wants to lead the conversation, what aspects of a topic interest him, how certain he is about what he is saying, how well he understands what his partner is saying, and so on. Such things can, in principle, be expressed explicitly, but people seldom do, and computers also should not, since in most situations such interpersonal and 'meta' aspects of communication must be dealt with deftly and concisely. One alternative is to express these things is by sensitive choice of words, such as the confirmations *un*, *hai*, *so*, *soso* and so on (Tsukahara 1998); another is careful manipulation of 'tone of voice'. Another technique is production of suitable grunts, primarily as back-channel feedback.

One advantage of grunts is that they may be easier for the hearer to process than full utterances. Grunts are phonetically simple, and tend to be more stable than ordinary utterances: a single spectral pattern persists for tens or hundreds of milliseconds. Whereas it is not generally possible to both talk and listen at the same time, it is possible to listen to grunts while talking. In this sense, grunts provide a separate channel, information in this channel does not much interfere with information in the verbal channel.

In cases where it does not suffice to simply select among pre-recorded grunts to play back, there arises the problem of generating grunts according to the meaning which needs to be expressed. For this some form of model-based synthesis for superimposing the appropriate acoustic components will be required, as simple concatenative techniques will be insufficient.

6 Applications for Grunt Understanding

Today it is common to provide information over the telephone with pre-recorded messages, such as weather reports and directions. These systems are frustrating for listeners, who have no control over the pace or content of the information, except via clumsy touch-tone commands. Recently there have been several proposals for systems that listen to how the user responds and adjust the playback rate or content accordingly (Iwase 1998). These systems use prosodic information or standard phoneme-based word recognition. By extracting the meaning of back-channel grunts from the user, using the correspondences outlined above, it should be possible to better understand and adapt output to the user's needs.

As noted above, grunts are acoustically different from words, in showing less temporal variation. This means that it may be possible to recognize grunts even when they overlap with playback by the system, even over telephone channels without perfect echo cancellation. This also means that grunt understanding should probably be done with spectral features computed over wider analysis windows than those used for word recognition, perhaps 50 or 100 milliseconds. The problem of grunt understanding is different form that of word recognition also in that there is no sequence of phonemes to recognize, but rather, the various strengths of many superimposed acoustic components must be computed. We are currently working on this problem.

The mapping from sound to meaning will not be invariant, but will depend on context. For example, if the previous utterance was *sugoi* 'wonderful', an *un* 'mm' will seem lukewarm, but if the previous 'utterance' was a silent stare, the same *un* can sound warm and friendly. Even in limited domains, where variations of context are not extreme, compensation for inter-speaker differences in the frequency of use of the various components of grunts will need to be taken into account.

Once the meaning of a grunt has been understood, the question of how the system should respond to it remains. One problem may be the integration of interaction at grunt-based (interpersonal and attitudinal) levels with interaction at the meaning level. It is possible to imagine implementing these as semi-independent response pathways, integrated with a subsumption architecture (Ward 1997).

References

Drummond, Kent & Robert Hopper (1993). Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency. *Research on Language and Social Interaction*, 26:157–177.

Gardner, Rod (1997). The Conversation Object *Mm*: A weak and variable acknowledging token. *Research in Language and Social Interaction*, 30:131–156.

Heritage, John (1984). A Change-of-State Token and Aspects of its Sequential Placement. In J. Maxwell Atkinson & John Heritage, editors, *Structure of Social Actions: Studies in Conversation Analysis*, pp. 299–345. Cambridge University Press.

Iwase, Tatsuya (1998). Yuza ni awaseta Taiwa Peesu no Chosetsu (Adjusting the Pace of Conversation to Suit the User). In *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pp. 472–475.

Kawamori, Masahito, Takeshi Kawabata, & Akira Shimazu (1995). A Phonological Study on Japanese Discourse Markers. In *9th Spoken Language Processing Workshop Notes (SIG-SLP-9)*, pp. 13–20. Information Processing Society of Japan.

Okada, Misao (1996). How the Length and Pitch of Aizuti 'Back-channel Utterances' and the Nature of the Speech Activity Determine Preference Structure in Japanese. In *Berkeley Linguistics Society, Proceedings of the Twenty-Second Annual Meeting*, pp. 279–289.

Shinozaki, Tubasa & Masanobu Abe (1997). Kisoku Gosei Onsei de Yakudokan o Jitsugen suru Horyaku ni tsuite (A Strategy for Realizing Live Interaction with Synthesized Speech). In *17th Spoken Language Processing Workshop Notes (SIG-SLP-17)*, pp. 81–88. Information Processing Society of Japan.

Takubo, Yukinori & Satoshi Kinsui (1997). Otoshi, Kandoshi no Danwateki Kino (The Conversation Functions of Responses and Exclamations). In *Bunpo to Onsei (Speech and Grammar)*, pp. 257–279. Kuroshio, Tokyo.

Tsukahara, Wataru (1998). Purosodi oyobi Bunmyaku Joho o Mochiita Ooto no Sentaku/Chosetsu no Kokoromi (Selecting and Adapting Confirmations in Response to Prosodic Indications and Contextual Factors). In *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pp. 468–471.

Ward, Nigel (1996). Using Prosodic Clues to Decide When to Produce Back-channel Utterances. In *International Conference on Spoken Language Processing*, pp. 1728–1731.

Ward, Nigel (1997). Responsiveness in Dialog and Priorities for Language Research. *Systems and Cybernetics*, 28(6):521–533.

Ward, Nigel & Wataru Tsukahara (1998). Prosodic Features which Cue Back-Channel Responses in English and Japanese. manuscript.

White, Sheida (1989). Backchannels across cultures: A study of Americans and Japanese. *Language in Society*, 18:59–76.