The Challenge of Non-lexical Speech Sounds

Nigel Ward[1]*
The University of Tokyo

ABSTRACT

Non-lexical speech sounds (conversational grunts), such as *uh-huh*, *un-hn*, *mm*, and *oh*, are common in English. In human dialogs these sounds are important in conversation control and for conveying attitudes. Spoken dialog systems may make use of these sounds to achieve concise, smooth, relaxed interactions. Doing so is, however, a challenge, because most algorithms used in spoken language processing were devised for words, but grunts are different from words both phonetically and semantically. For example, the phonetic inventory is different, superimposition of phonemes occurs, the set of conversational grunts is productive rather than finite, and the meanings are compositional and involve sound-symbolism.

1 PURPOSE

This paper discusses non-lexical speech sounds ('conversational grunts' for short) in dialog, and points out possibilities and problems for their use in spoken language processing.

2 GRUNTS IN A CORPUS OF CONVERSATIONS

This section presents some facts about the frequency and uses of grunts in human-human conversation.

Non-lexical conversational sounds, such as *uh-huh*, *un-hn*, *mm*, and *oh*, are ubiquitous in informal spoken English. In our data, these grunts occur an average of once every 5 seconds in American English conversation. In a sample of conversations from Switchboard, *um* was the 6th most frequent item (after *I*, *and*, *the*, *you*, and *a*), and the four items *uh*, *uh-huh* and *um* and *um-hum* accounted for 4% of the total (Picone *et al.* 1998).

Tables 1 and 4 illustrate the diversity of phonetic forms and functional roles taken by conversational grunts. Most studies of conversational grunts have focused on one specific functional role, such as filler, back-channel or disfluency marker. However, from Table 1 it is clear that many items occur in a variety of functional roles, suggesting that it may

	ttl.	back	fill	disfl	is	rs	c	o
[click]	22	.	12	2	1	.	.	7
ah	7	1	3	3	.	.	.	.
aum	5	.	4	1	.	.	.	.
hh	3	.	.	.	2	.	.	1
mmm	3	2	1	.	.	.	.	.
nn-hn	4	4	.	.	.	.	.	.
oh	20	6	9	.	.	.	.	5
okay	8	2	2	.	.	1	2	1
u-uh	4	.	.	2	.	2	.	.
uh	38	.	14	21	1	.	.	2
uh-huh	3	3	.	.	.	.	.	.
um	20	.	10	8	.	.	.	2
umm	5	.	5	.	.	.	.	.
uu	5	2	2	.	.	.	.	1
uum	5	.	3	2	.	.	.	.
yeah	71	27	19	1	6	6	6	6
(other)	94	44	24	5	10	4	.	7
Total	319	91	110	45	20	13	8	32

Table 1: Counts of Grunt Occurrences in various positions and functional roles, for all grunts occurring 3 or more times in our data. Labeling criteria are given elsewhere (Ward 2000a). Functional roles and orthographic conventions are described in Tables 2 and 3.

abbreviation	function/position
back	back-channel
fill	filler, including various things that occur utterance- or turn- initially
disfl	disfluency marker
is	isolate, produced when neither person has the turn, typically more self-directed than other-directed
rs	response to direct question or high-rise statment
c	confirmation, in response to a back-channel
o	other, including clause-final items, items that occur in quotations, and items whose function is obscure

Table 2: Functional/Positional Roles of Conversational Grunts. Details appear in (Ward 2000b).

*Thanks to the Sound Technology Promotion Foundation, the Nakayama Foundation, the Inamori Foundation, the International Communication Foundation, and the Japanese Ministry of Education for support.

[1] nigel@sanpo.t.u-tokyo.ac.jp Mech-Info Engineering, University of Tokyo, Bunkyo-ku Tokyo 113-8656 Japan. http://www.sanpo.t.u-tokyo.ac.jp/∼nigel/

[clear-throat]	2	hh	3	nu	1	u-uun	1	uuh	1
[click]	22	hh-aaaah	1	nuuuuu	1	uam	1	uum	5
[click]naa	1	hhh	1	nyaa-haao	1	uh	38	uumm	1
[click]neeu	1	hhh-uuuh	1	nyeah	1	uh-hn	2	uun	1
[click]ohh	1	hhn	1	o-w	1	uh-hn-uh-hn	1	uuuh	1
[click]yeah	1	hmm	2	oa	1	uh-huh	3	uuuuuuu	1
[inhale]	1	hmmmmm	1	oh	20	uh-mm	1	wow	1
aa	1	hn	1	oh-eh	1	uh-uh	2	yah-yeah	1
achh	1	hn-hn	1	oh-kay	1	uh-uhmmm	1	ye	1
ah	7	huh	2	oh-okay	2	uhh	2	yeah	71
ahh	1	i	1	oh-yeah	1	uhhh	1	yeah-okay	1
ai	1	iiyeah	1	okay	8	uhhm	1	yeah-yeah	1
am	1	m-hm	2	okay-hh	1	ukay	2	yeahaah	1
ao	1	mm	2	ooa	1	um	20	yeahh	1
aoo	1	mm-hm	1	ookay	1	um-hm-uh-hm	1	yegh	1
aum	5	mm-mm	1	oooh	1	umm	5	yeh-yeah	1
eah	1	mmm	3	ooooh	1	ummum	1	yei	1
ehh	1	myeah	2	oop-ep-oop	1	unkay	1	yo	1
h-nmm	1	nn-hn	4	u-kay	1	unununu	1	yyeah	1
haah	1	nn-nnn	1	u-uh	4	uu	5		

Table 4: All Grunts in our Data, with numbers of occurrences. All of these items appear to be different in meaning. The orthographic conventions are described in Table 3.

be profitable to treat the set of conversational grunts as a whole, which is the approach taken in this paper.

The pragmatic functions borne by these sounds relate primarily to attitudinal dimensions of the interaction (where a participant indicates how pleased he is with the current topic, how interested he is, how well he understands it, and so on) and to conversation control functions (where a participant indicates whether he wants to lead the conversation, whether he finds the pace too fast or too slow, and so on).

3 POTENTIAL APPLICATIONS

Since conversational grunts are thus common and useful in human communication, they should have value in spoken language processing also. This section gives a few examples.

1. Conversational Grunt Understanding in Dialog Systems. Systems which provide information over the telephone, such as weather reports and directions, can be frustrating for listeners, who typically have no control over the pace or content, except via clumsy touch-tone commands or spoken equivalents such as *repeat* and *main-menu*. Instead systems probably should allow users to control information transmittal by human-like protocols (an idea that has been around at least since Schmandt (Iwase & Ward 1998)). Various demonstrations have shown that fairly natural exchanges can result even if a system only uses gross information from a user's grunts, such as their presence/absence, length, and prosody, and uses this to decide whether to wait, repeat, or go on. It should be possible to allow finer-grained control by recognizing the content of user grunts, such as *uh-huh* meaning "go on, don't talk so slow", *uh-hum* meaning "slow down, I need to think", and *ah* meaning "I have something to say". More generally, communication using conversational grunts may be preferable to full sentences as a concise and informal way to handle the attitudinal and 'meta' aspects of interaction, which are important parts in all but the most formalized interactions.

2. Grunt-based Search. In speech databases, users may need to search for information that is sometimes conveyed by grunts, such as skepticism, amusement, decisiveness, the presence of important information, and so on.

3. Transcriptions including Grunts. Automatic transcriptions of dialogs might be more useful if they include grunts and annotations, such as *uh-huh (non-committal)* or *uh-huhh (amused)*.

4. Producing Grunts for Acknowledgements etc. The fact that grunts are short makes them potentially valuable as prompts, acknowledgments, and confirmations. Moreover, their informal nature may allow the construction of systems with more casual, friendly personalities. Furthermore, grunts may be easier for hearers to process than full utterances: whereas humans can not generally both talk and listen at the same time, they can listen to grunts while talking — in this sense, grunts provide a separate channel, where the use of this channel does not much interfere with the main channel.

For these reasons, grunts as system output may support much swifter turn-taking. This is illustrated by a recent experiment involving a simple memory game (Tsukahara & Ward 2000). The game starts like this: "can you name all 29 stations of the Yamate loop line? Say them in or-

notation	phonetic value
h	a single syllable-final 'h' bears no phonetic value, elsewhere 'h' indicates /h/ or breathiness
n	nasalization
click or tsk	alveolar tongue click
gh	velar fricative
chh	palatal fricative
u	schwa
uu	as a syllable, indicates a short creaky or glottalized schwa
oop	/up/
repetition of a letter	length and/or multiple weakly- separated syllables
- (hyphen)	a fairly strong boundary between syllables or words (typically realized as a major dip in energy level, a sharp discontinuity in pitch, or a region of breathy or creaky voice)

Table 3: Orthographic Conventions (non-obvious aspects)

der, and I'll give you hints if you get stuck". Although this task is semantically very limited, it can be entertaining. By using grunts for acknowledgements (roughly the Japanese equivalents of *yeah*, *uh-huh*, *mm*, *mm-hm*, *yeah*, *okay* and *right*), it was possible to produce a system able to keep up the same swift pace as an exemplary human tutor (allowing dialog to continue at a cycle time of as little as 1.6 seconds from one guess to the next, including the intervening acknowledgement from the system), enabling users to get completely involved in the game of recalling as many station names as possible in the time allotted. Moreover, even at this pace users were sensitive to the specific grunts used, rating more highly the system which chose grunts in order to praise, encourage, express pleasure, back off to give the user more time, and so on, as appropriate for each situation.

5. Producing Fillers and Disfluency Markers. Different users have different information-uptake capabilities. Inserting fillers and disfluencies in system output may be a relatively easy way to reduce the information transmission rate. Appropriate fillers and disfluency markers may also assist the listener by signaling what sort of information is coming up, how long it will be, and so on, so that he can deploy his attention appropriately.

6. Detecting Grunts. Today the primary application for conversational grunts is detecting them so they can be ignored. This is true for fillers and disfluency markers since they interfere with recognition of neighboring words. It may also be true for back-channels in some telephony applications, where they may obscure the speech of the other party.

sound	items incl.	meaning
schwa	109	neutral
/j/	82	solid understanding?
syllabification	57+	lack of anything to add
/m/	56	contemplation
creaky voice	53	detachment
/o/	45	new information
/h/ and breathiness	38	engagement
clicks	25	dissatisfaction
nasalization	20	shared knowledge
/a/	5	readiness to act

Table 5: Common and Salient Phonetic Components of Grunts, the total number of grunts which include each component, and hypotheses regarding meanings (highly abbreviated).

4 PHONETIC AND SEMANTIC PROPERTIES

The development of such applications is complicated by the fact that conversational grunts differ from words, in many ways. This section summarizes the differences, omitting references and evidence for lack of space.

P1. Unusual Phonetic Segments. Conversational grunts involve acoustic components outside the normal phonology of the language, including clicks, nasal vowels, and glottal stops. While the more exotic sounds, such as uvular fricatives, are rare, others are common even in such mundane roles as back-channels.

P2. Unusual Voicing. Conversational grunts often include significant breathiness and creakiness.

P3. Limited Phonetic Inventory. The inventory of sounds found in conversational grunts is fairly limited, excluding most of the phonemes present in lexical items, including high vowels, plosives, and most fricatives.

P4. Superimposition of Phonetic Components. Conversational grunts are generated not only by concatenation of phonetic components but also by superimposition, for example, in a segment that is simultaneously nasal and creaky and central and low-pitched.

P5. Spectral Stability. Conversational grunts tend to be more stable than ordinary utterances: a single spectral pattern can persist for hundreds of milliseconds.

P6. Productivity. Many conversational grunts appear to be created on the fly, according to the communicative needs of the speaker, rather than simply being selected from a finite set of fixed phoneme strings (although there is a gradient, from items with limited variation, like *okay* and *yeah*, to completely malleable items, such as *uh-huh*, *uh-hn* and *um-hm*).

P7. Sound Symbolism. Each of the component sounds seems to bear some meaning or function (which implies that observations P1, P2, and P4 are indeed significant.) Moreover, it seems that these meanings are fairly constant across grunts and across contexts, as seen in Table 5, and thus the meanings of conversational grunts are largely compositional, or, in other words, involve sound symbolism.

P8. Non-Categorical Phonology. Some of the acoustic components of grunts, such as degree of nasalization, degree of breathiness and pitch height, seem to be present to a greater or lesser degree, with this degree conveying meaning, rather than being categorical (simply present or absent). For example, the difference between an *uh-huh* of agreement and an *uh-uh* of denial is probably dependent on several non-binary features, including degree of breathiness, syllable boundary strength, and pitch slope angle in the second syllable.

P9. Context-Dependent Functions. The pragmatic force borne by any specific grunt depends not only on its intrinsic, compositional meaning but also on the discourse context, in complex ways.

P10. Limited Syntactic Affinities and Collocational Tendencies. The occurrences of the various grunts seem to be relatively unpredictable from local context, compared to words.

P11. Significant Prosody. Much of the information in conversational grunts is borne by their prosody. However the prosody of grunts seems to be relatively simple (the most meaningful prosodic features are probably: loudness, pitch height, pitch slope, number of syllables and degree of syllabification, duration, and abruptness of final energy drop).

P12. Non-Propositional Meanings. The meanings and functions of conversational grunts are hard to formalize.

5 CHALLENGES

These properties of conversational grunts may cause difficulties for standard algorithms and techniques. More positively, developing algorithms which exploit the specific properties of grunts may give better performance. This section lists some issues.

For acoustic models P1∼P6 and P8 are relevant. These are also relevant for classifiers which discriminate conversational grunts from words, and indeed some of these properties have been already been exploited (Shriberg 1999; Goto *et al.* 1999).

For prosodic processing, the relative simplicity of grunts (P11) may make them fairly easy to deal with.

For language modeling, P10 suggests that discourse models may be more useful than models relying on local context. P6 and P7 may make it possible to model grunts with relatively few parameters, since the items can perhaps be predicted via predictions of the component sounds.

For dialog management, P9 makes dealing with grunts difficult.

For synthesis algorithms P1∼P6 and P8 are relevant.

For meaning representation and reasoning, the naive approach of treating each grunt item as a separate lexical item, with its own meaning, is unparsimonious, due to P7.

6 PROSPECTS

Clearly there is need for more basic linguistic research in conversational grunts. Preliminary study of Japanese suggests that properties P1∼P12 are also true in conversational grunts in that language (Ward 1998), but again more work is needed. The least understood aspect of grunts, but the most important for most applications, is the details of the sound-meaning correspondences, a problem we are currently working on. There is also a need for research in the specific technical issues mentioned above, to develop methods suited to the specific properties of grunts. We also need to build systems to determine the utility of grunts in applications.

7 REFERENCES

Goto, Masataka, Katunobu Itou, & Satoru Hayamizu (1999). A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. In *Eurospeech '99*, pp. 227–230.

Iwase, Tatsuya & Nigel Ward (1998). Pacing Spoken Directions to Suit the Listener. In *International Conference on Spoken Language Processing*, pp. 1203–1206.

Picone, Joe *et al.* (1998). Switchboard Statistics: Word Statistics. Institute for Signal and Information Processing, Mississippi State University, http://www.isip.msstate.edu/projects/switchboard/.

Shriberg, Elizabeth E. (1999). Phonetic Consequences of Speech Disfluency. In *Proceedings of the International Congress of the Phonetic Sciences, Volume 1*, pp. 619–622.

Tsukahara, Wataru & Nigel Ward (2000). Evaluating Responsiveness in Spoken Dialog Systems. In *International Conference on Spoken Language Processing.*

Ward, Nigel (1998). The Relationship between Sound and Meaning in Japanese Back-channel Grunts. In *Proceedings of the 4th Annual Meeting of the (Japanese) Association for Natural Language Processing*, pp. 464–467.

Ward, Nigel (2000a). Issues in the Transcription of English Conversational Grunts. submitted to the First (ACL) SIGdial Workshop on Discourse and Dialog.

Ward, Nigel (2000b). Labelers Manual for the Converational Grunt Project, Version 2.1. http://www.sanpo.t.u-tokyo.ac.jp/∼nigel/manual.html.