Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Pragmatic Functions of Prosodic Features in Non-Lexical Utterances

Nigel Ward

Department of Computer Science
University of Texas at El Paso
nigel@cs.utep.edu

Abstract

In informal English dialog many utterances are not composed of words, but are non-lexical items, such as uh-huh, um, and hmm. In non-lexical utterances much of the meaning is conveyed by prosody, rather than by the phonetic content. However the pragmatic functions of prosody in non-lexical utterances have not been much studied. Based on examination of 316 tokens in a conversation corpus, this paper identifies some common pragmatic functions for syllabification, duration, loudness, pitch height, pitch slope, and creaky voice in non-lexical utterances. While the evidence is eclectic and the investigation has been unsystematic, it seems that each of these prosodic features bears a fairly consistent core meaning.

10.21437/SpeechProsody.2004-76

1. Introduction

Prosody has been studied largely in utterances composed of words. However informal human communication also makes heavy use of non-lexical utterances, such as uh-huh, oh, and umm. Prosody is especially important in such items, where it often conveys more meaning than does the phonetic content. Indeed Bolinger notes that sometimes such an item 'might almost be regarded as a mere intonation carrier' [1].

This is frequently seen in dialog, where non-lexical utterances functioning as back-channels, fillers, disfluency markers and the like rely heavily on prosody to perform their functions, which include turn-taking control, negotiating agreement, signaling recognition and comprehension, managing interpersonal relations such as control and affiliation, and expressing emotion, attitude, and affect.

This paper is an attempt to remedy the lack of attention to the relationships between prosody and pragmatics in non-lexical utterances. The aim is to identify the most common meanings for the most commonly occurring prosodic features.

2. Methods

Studies of dialog phenomena most commonly rely on either a detailed study of a handful of examples or a statistical analysis of a large corpus. However, given the current state of knowledge, it was felt that a moderately thorough study of a few hundred examples would be the most productive approach [2]. Thus this study examined a small corpus of casual American English conversations [3]. The aim was to exhaustively describe all 316 non-lexical utterances in the corpus, and from that to create a formal model of the relationship between sound and meaning. Needless to say, this is not yet complete; this paper reports the findings to date.

This corpus includes, for each non-lexical item, labels for 8 dimensions of pragmatic function. Although these dimensions were selected for another purpose, namely examination of the pragmatic functions associated with various phonetic features [2], they were sometimes were useful here too. These labels were generated by 2 native-English speaking judges not including the author. In this corpus all non-lexical items were also phonetically labeled by the author and an advanced phonetics student.

Working with this corpus, hypotheses regarding the meaning of each pragmatic feature were generated. This was done by repeatedly listening to the various non-lexical utterances in context, in order to iteratively approach the meaning/function description which best accounts for the meanings of almost all the occurrences in the corpus without being overly general. Given the goal of identifying meanings for prosodic featuers, several working assumptions were adopted, including, first, that the meanings of the prosodic features are compositional and thus the contribution of each is evident in the meaning of the whole; second, that the meanings of the prosodic features are orthogonal to and not affected by the functional position (filler vs. back-channel etc.) in which the item appeared; and third, that the meanings of the prosodic features are orthogonal to and not affected by the phonetic content of the non-lexical items. Although clearly not always valid [2], in practice these assumptions were generally unproblematic. It was further assumed that the prosodic features are continuous, not categorical, and that subjective judgments are accurate: these assumptions made the analysis possible.

After the hypotheses were generated they were evaluated against the corpus. This was done opportunistically rather than systematically, using whatever information in the labels or distributions could be brought to bear. The analysis also included examination of minimal pairs or near minimal pairs, ideally differing only the strength or presence/absence of one prosodic feature. Some of these minimal pairs appear below as illustrations of the meanings involved; pitch diagrams and audio for these are available at <http://www.cs.utep.edu/egrunts/>.

This work was supported in part by the International Communications Foundation and by the Japanese Ministry of Education's Prosody and Speech Processing Project, headed by Keikichi Hirose.

Example 1: discussing a party they might go to

H: Is it like a party, like, 'rave' type party? or like	1
C: well, it's someone's house	2
H: yeah	3
C: there's going to be, I mean there's like, they're going to be spinning. So, in that sense, maybe, but it's just at someone's house, like	4
H: yeh-yeah	5
C: it's in the middle of the night, that too, but.	6

Example 2: T is driving, O is navigating

O: can we turn here? can, can we make a right turn here?	1
T: If you say so	2
O: um, oh, I guess we can't (embarrassed laugh). No. (laugh)	3
T: what? no.	4
O: uuuh. hmm	5
T: should we turn around and go back?	6
O: uh-mm ...(waits until the next intersection comes up before deciding)	7

3. The Features and their Functions

3.1. Syllabification

Syllabification is a very salient property of non-lexical items. Unlike other prosodic features, syllabification is even reflected in the conventional spellings, as in mm-mm vs. mm, uh-huh vs. uh and yeah-yeah vs. yeah.

Two-syllable items often signal the intention to take a listening role, to indicate that the person who produces them intends to say no more. Evidence for this includes the fact that yeah-yeah only functions as a back-channel, in contrast to yeah which appears in many roles. Similarly uh-huh and um-hm are overwhelmingly back-channels, versus single-syllable uh and um which are overwhelmingly fillers and disfluency markers.

One speaker produced four-syllable items, uhn-hm-uh-hm and um-hm-uh-hm, and these appeared to contrast with um-hm: the four-syllable forms signaled a posture of continued listening, but the two-syllable um-hm was less passive, sometimes produced only shortly before he interrupted and took a turn.

By implication, the fact that you have nothing to add can serve to be encouraging the interlocutor to continue. Often, as with uh-huh, this is a purely passive posture. Other times, as with yeah-yeah, this can encourage the interlocutor to stop repeating himself and get to the point, as in Example 1 line 5. (Incidentally, yeah-yeah in a creaky voice, and with a sharp downstep in pitch to add brusqueness, is a stereotypical way to say 'enough already, let's drop this topic'.)

Although multiple syllables occur most commonly in back-channels, some syllabification also occurs in other positions, and with the same meaning. In Example 2 line 5 the uuuh has three energy peaks, and sounds frustrated: this can be ascribed to the fact that O wanted to say what to do next (for the sound appears where it can only be interpreted as a filler), but is simultaneously realizing that he doesn't know and so can say no more, as conveyed by the syllabification.

Example 3: F is starting to explain his research

F: click, inhale, trying to to develop, models of um (1.5 second pause) uh word models, word phonological models that sort of, match acoustic data better and, is able to be, modified by context ...	1

Example 4: F is continuing to explain his research

F: the phone recognition and then learn the, transformation, between the two streams, and, uh a second level thing was to, then, after we've built this transformation, automatically learn that transformation for each phoneme, given, a particular acoustic context given, ...	1

In general, syllabification in a non-lexical token seems to disclaim the intention to say anything more, to indicate that the producer is for the moment content to listen and/or remain silent. Excluding intrinsically multi-syllabic items such as okay, of the tokens with syllabification, 63% (38 of 60) seem to be indicating such a lack of anything to say, compared with about 4% for single-syllable tokens.

It is worth noting that multi-syllable tokens are generally not simply repetitions of a single syllable. Rather they generally include one or more additional features marking the syllable boundaries, most commonly energy dip, pitch dip, breathiness, or creakiness, and these occur at various strengths. Thus the term 'syllabification' is more appropriate than 'reduplication'. The choice of how to realize syllabification is perhaps independent of the choice of syllabification itself; thus, for example, when a syllable boundary marked with breathiness is present, it may convey both the meaning of breathiness [2] and the meaning of syllabification.

3.2. Duration

It is well known that the duration of a filler before an answer correlates with uncertainty regarding the response. More generally, the longer the filler, the more the person is considering what he plans to say. In example 3 there is a very short (100ms) uh which appears when F has apparently figured out what he wants to convey (subsequent delivery is fairly fluent), but is just trying to chose the correct phrase. In contrast, the uh in example 4 lasts 950ms, and occurs where F is struggling with a complicated new topic.

Something similar is true for back-channels. For example, the sympathetic mmm in Example 5 line 6 lasted 580 milliseconds, but the mm in Example 6 line 4 a mere 360 milliseconds, as is appropriate for a lighter topic.

In general, duration appears to correlate with thought, where the thought relevant in dialog includes both thought involved in speaking and thought involved in listening. This correlation was evaluated in several ways.

First, the corpus contained labels indicating which non-lexical utterances seemed to express "deepness". While this is not the same as "thoughtfulness", most cases of thoughtfulness probably involve deepness, and conversely. Here, as so often with this sort of pragmatic function, inter-labeler agreement was low. Limiting attention to the 14 tokens which both labeled "deep", the

Example 5: after some talk about television, children, and violent play

X:	and this video was about Ultraman . . . most of it's not too violent . . . but there is a little bit of stabbing and stuff	1
M:	right	2
X:	and so he came home and he was stabbing poor little Henry	3
M:	nyaa-haao	4
X:	yeah, I, I felt.	5
M:	mmm	6
X:	well, I mean, yeah. .click. I was pretty annoyed.	7

Example 6: after some generalities about what sort of people read Japanese comics

N:	There's one student, he's got his desk, and his bookcase, and his bookcase is filled with, well, books, but	1
M:	right	2
N:	most of them are comics	3
M:	mm	4
N:	and on the top he has a row of Sailor Moon dolls . . .	5

average duration of was 486 ms, longer than the average of the others, 365 ms. The distribution of duration of these deep tokens was significantly different from the overall distribution ($p < .024$, one-tailed t-test, assuming normal distributions).

Second, the author labeled all non-lexical utterances on a five-point scale, as seen in Table 1. Tokens involving more thought were generally longer, with the differences between t0 and t1, t2 and t3, and t3 and t4 significant by t-test. Similar correlations between duration and the t0-t4 scale were seen within various functional types: disfluencies, fillers, back-channels, and also when limiting attention to tokens of yeah.

Third, the relation between duration and thought may also be seen in the average durations of non-lexical items across various functional roles: the average duration of disfluencies being 313 ms, fillers 328 ms, and back-channels 415 ms.

It is interesting that these correlations show up even when duration is measured crudely, as here, without normalization to local speech rate or adjustment for cases where an inbreath close to a non-lexical utterances may have affected the perceived length.

There were some exceptions to the duration-thought correlation. A few long tokens did not seem to involve

degree of thought	n	average duration
t0: no thought (reflex responses, etc.)	81	289 ms
t1: mild thought	96	333 ms
t2: some thought	54	364 ms
t3: a lot of thought	38	576 ms
t4: intense thought	14	768 ms
tx: impossible to determine	28	
all	310	376 ms

Table 1: Relation between Thought and Duration

Example 7: C has applied for a summer-abroad program

H:	I bet you'll hear something soon.	1
C:	I hope so. I just turned that in, though, like. A couple weeks ago, so.	2
H:	yeah (slightly creaky)	3
C:	you know what I mean, so	4
H:	yeah, it might take a little longer	5
C:	nn-hn	6

Example 8: 30 seconds later H discusses her own application to a summer-abroad program

H:	well, I got an e-mail, that said that I was, like recommended	1
C:	uh-hn	2
H:	and it said, it was like, to me and one other girl	3

thought, but rather impatience, or politeness, or pacing control. Also there were a few clicks, necessarily short of course, which somehow did seem to be thoughtful.

Incidentally, there is little or no correlation between thoughfulness and the number of phonemes: thus this duration effect appears to be a stretching out of some phonetic content.

3.3. Height

In general, pitch height seems to correlate with degree of interest.

Comparing the nn-hn in Example 7 with the uh-hn in Example 8, the former is not only quieter, but also lower pitch in both the first syllable and the second. It ended this topic of conversation. The second, higher-pitched token shows more interest, and here the topic was continued.

Similarly in Example 1, the second token of yeah, in addition to being bisyllabic, is of lower pitch. Often topics seem to exhibit a sort of life cycle in which the back-channels start high and go down as the topic winds down. There is a concomitant tendency for the back-channels to get quieter and the downslope to get weaker.

Overall non-lexical items have a weak tendency to be lower in pitch than do words. Fillers however have a tendency to be higher, perhaps related to the fact that grabbing the turn is a common use of high pitch. Disfluency markers, in contrast, are overwhelmingly low in pitch.

3.4. Loudness

While a general analysis of the significance of volume has not yet been done, there is one salient phenomenon in the corpus: the existence of corpus items which were perceptually very quiet. Most of these were in back-channel positions, and many were sounds without vowels, such as mm and hh. Some of these seemed so quiet that, although picked up by the head-mounted microphones, they were probably not perceptible to the conversation partner. It is hard to imagine any pragmatic function being served by such utterances. Rather they may be useful for the study of real-time cognitive processing as it relates to dialog.

In general one would expect, common-sensically, loudness to correlate with assertiveness, self-confidence, and the importance of the utterance.

3.5. Pitch Slope

Systematic study of pitch slope has also not yet been done. However a preliminary line-fitting exercise showed, contrary to expectation, that the vast majority of non-lexical utterances have a very flat pitch, even in relatively long tokens. Of course it is well known that flat pitch is a distinguishing characteristic of fillers and disfluency markers [4], but this is seen in back-channels also.

Of the remainder, most have a simple falling pitch, and they seem to convey something like decisiveness. A tiny number have a rising pitch, and these seem to function mostly as questions or challenges.

3.6. Pitch Contours

One initial motivation for this study was the existence of non-lexical utterances with complex pitch contours and complex meanings, such as the contours signifying "yes, definitely", "yes it is", "no it isn't", "I'm disappointed", "no way", etc., noted by Luthy [5] and Ehlich [6], among others. Unfortunately, in this corpus complex contours were vanishingly rare, perhaps because all the conversations were seated interactions between polite adults, so the following comments are speculative.

Most of the complex contours occurr on the multi-syllable tokens. A classic example is uh-huh, where the first syllable stereotypically has a flat or falling pitch, and the second a rising or flat pitch. These may represent two dialog acts which, being temporally adjacent, blend into one utterance.

There was one token which seemed to bear sentence-like prosody. Pragmatically, this seemed to be substituting for a full turn in the main channel. There are a few cases where pitch contours are falling but curved rather than linear; these may be Californianisms. Occasionally there are small pitch upturns at the end of a token.

3.7. Other Prosodic Factors

Creaky voice, marginally a prosodic feature, appears to encode a variety of meanings, but most often to convey a sort of authority. Although people sometimes say things lightly, other times they really know what they are talking about. Thus some things people say in conversation are intended as authoritative statements: advice, opinions, decisions, recollections, etc., based on expert knowledge or direct experience. Creaky non-lexical utterances generally convey such a meaning. The evidence is given in [2].

The timing of non-lexical utterances, relative to other utterances by the same speaker or relative to those of the dialog partner, is important and deserving of systematic study.

There are probably also other meaningful prosodic features. For example, abruptness of energy drop, giving a clipped sound, may be a 'gesture of finality' [7].

4. Summary and Discussion

Table 2 summarizes the meanings tentatively attributed to each prosodic feature. None of the meanings found for the prosodic features is particularly surprising; rather all are pretty much in line with what is seen in lexical utterances. This was contrary to the author's expectation.

sound	meaning
syllabification	lack of desire to talk
duration	amount of thought
pitch height	degree of interest
loudness	confidence, importance
pitch downslope/upslope	degree of understanding / lack thereof
creaky voice	assertion of authority

Table 2: Summary

Also unexpected was the paucity of complex pitch contours.

Intriguingly, some of these prosody-meaning correlations also appear in Japanese non-lexical items, as seen in an analogous corpus of Japanese conversations [8], although probably not across all speaking styles [9].

While there are special cases and sub-generalizations, not discussed here for lack of space, the major tendencies are consistent across the data. The prosodic functions identified here are necessarily vague, but for building specific applications [10], it should be easy to refine them into more precise, dialog-type-specific meanings.

5. References

[1] Bolinger, Dwight, 1989. Intonation and Its Uses. Stanford University Press.

[2] Ward, Nigel, 2003. Non-lexical Conversational Sounds in American English. manuscript.

[3] Ward, Nigel and Wataru Tsukahara, 2000. Prosodic Features which Cue Back-Channel Feedback in English and Japanese. J. of Pragmatics, 32, 1177–1207

[4] Shriberg, Elizabeth, 2001. To 'errr' is Human: Ecology and acoustics of speech disfluencies. J. of the International Phonetic Association, 31, pp 153–169.

[5] Luthy, Melvin J., 1983. Nonnative Speakers' Perceptions of English "Nonlexical" Intonation Signals. Language Learning, 33, pp 19–36.

[6] Ehlich, Konrad, 1986. Interjektionen. Tuebingen: Max Niemeyer Verlag.

[7] Bolinger, Dwight, 1946. Thoughts on 'Yep' and 'Nope'. American Speech, 21, pp 90–95.

[8] Ward, Nigel, 1998. The Relationship between Sound and Meaning in Japanese Back-channel Grunts. Proc. of the 4th Ann. Meeting of the (Japanese) Ass'n for Natural Language Processing, pp 464–467.

[9] Iida, Akemi, Parham Mokhtari and Nick Campbell, 2003. Acoustic Correlates of Monosyllabic Utterances of Japanese in Different Speaking Styles. Int'l Congress of the Phonetic Sciences, pp 2861-2864.

[10] Ward, Nigel, 2000. The Challenge of Non-lexical Speech Sounds. Int'l Conference on Spoken Language Processing, II: 571–574.