# Issues in the Transcription of English Conversational Grunts

**Nigel Ward**
Mech-Info Engineering, University of Tokyo,
Bunkyo-ku, Tokyo 113-8656, Japan
nigel@sanpo.t.u-tokyo.ac.jp
http://www.sanpo.t.u-tokyo.ac.jp/~ nigel/

## Abstract

Conversational grunts, such as *uh-huh*, *un-hn*, *mm*, and *oh* are ubiquitous in spoken English, but no satisfactory scheme for transcribing these items exists. This paper describes previous approaches, presents some facts about the phonetics of grunts, proposes a transcription scheme, and evaluates its accuracy.[1]

## 1 The Importance of Conversational Grunts

Conversational grunts, such as *uh-huh*, *un-hn*, *mm*, and *oh* are ubiquitous in spoken English. In our conversation data, these grunts occur an average of once every 5 seconds in American English conversation. In a sample of 79 conversations from a larger corpus, Switchboard, *um* was the 6th most frequent item (after *I*, *and*, *the*, *you*, and *a*), and the four items *uh*, *uh-huh*, *um* and *um-hum* accounted for 4% of the total. These sounds are not only frequent, they are important in language use. To mention just one example, people learning English as a second language are handicapped in informal interactions if they cannot produce and recognize these sounds.

[1] I would like to thank Takeki Kamiyama for phonetic label cross-checking, all those who let me record their conversations, and the anonymous referees; and also the Japanese Ministry of Education, the Sound Technology Promotion Foundation, the Nakayama Foundation, the Inamori Foundation, the International Communications Foundation and the Okawa Foundation for support.

Just to be clear about definitions, in this paper 'grunts[2]' means sounds which are 'not words', where a prototypical "word" is a sound having 1. a clear meaning, 2. the ability to participate in syntactic constructions, and 3. a phonotactically normal pronunciation. For example, *uh-huh* is a grunt since it has no referential meaning, has no syntactic affinities, and has salient breathiness. In this paper 'conversational' refers to sounds which occur in conversation and are at least in part directed at the interlocutor, rather than being purely self-directed[3]. Both of these definitions have flaws, but they provide a fairly objective criterion for delimiting the set of items which any transcription scheme should be able to handle.

The phenomena circumscribed by this definition are a subset of "vocal segregates" (Trager, 1958) and of "interjections": the difference is that it limits attention to sounds occurring in conversations. This definition also roughly delimits the subset of "discourse markers" or "discourse particles" which occur in informal spoken discourse.

As the phonetics and meanings of conversational grunts are currently not well understood, we have begun a project aiming to elucidate, model, and eventually exploit them. The current paper is a report on an approach

[2] It may seem that the negative connotations of the word 'grunt' make it inappropriate for use as a technical term, but the phenomenon itself is often stigmatized, and so the term is appropriate in that sense too.

[3] Two rules of thumb were adopted to help in cases which were difficult to judge: consider laughter as not conversational, and consider as conversational everything else that might possibly be playing some communicative role, even if it isn't clear what that role might be.

to the preliminary problem of how to transcribe these sounds.

A generally usable, standardized transcription scheme would be of great value. Immediate applications include screenplay writing and court recording. It would also facilitate the systematic corpus-based study of the meanings and functions of these sounds[4]. There are also prospects for applications in systems. One could imagine a dialog transcription system that produces output with the grunts represented in enough detail to show whether a listener is being enthusiastic, reluctant, non-committal, bored, etc., as these states are often indicated by grunts rather than by words. One could imagine spoken dialog systems which prompt and confirm concisely with such grunts, instead of full words or phrases. And one could imagine spoken dialog systems which adjust their output based on barge-in feedback from the user such as *uh-huh* meaning "go on, don't talk so slow", *uh-hum* meaning "stop, I need to think", and *ah* meaning "I have something to say".

Section 2 surveys previous approaches to grunt transcription, Section 3 proposes a slightly new scheme, Section 4 discusses its adequacy, and Section 5 points out some open issues.

## 2 Previous Schemes for Grunt Transcription

This section points out the problems with previous approaches to grunt translation.

### 2.1 Phonetically Accurate Schemes

One tradition in labeling grunts is to use a completely general scheme. The central inspiration here is the fact that grunts are unlike words, in that they contain sounds which are never seen in the lexical items of the language. As such, they can fall outside the coverage of even the International Phonetic Alphabet, which is only designed to handle those sounds which occur contrastively in some words in some language. Thus there have been proposals for richer, more complete transcription schemes, capable of handling just about any communicative noise that people have been observed to produce, including moans, cries and belches (Trager, 1958; Poyatos, 1975).

One disadvantage of these notations is that they are not usable without training.

A second disadvantage is that their generality is excessive for everyday use. As seen below, the vast majority of conversational grunts are drawn from a much smaller inventory of sounds.

A third disadvantage is that they provide more accuracy than is needed. For example, in English there appear to be no grunts in which the difference between an alveolar nasal, a velar nasal, or nasalization of a vowel conveys a difference in meaning, and so these do not need to be distinguished in transcription.

### 2.2 A Function-based Schemes

An alternative approach is seen in some schemes used for labeling corpora for purposes of training and evaluating speech recognizers. A quote from the most recent Switchboard labeling standard (Hamaker et al., 1998) gives the flavor:

> 20. Hesitation Sounds: Use "uh" or "ah" for hesitations consisting of a vowel sound, and "um" or "hm" for hesitations with a nasal sound, depending upon which transcription the actual sound is closest to. Use "huh" for aspirated version of the hesitation as in "huh? <other speaker responds> um ok, I see your point."
>
> 21: yes/no sounds: Use "uh-huh" or "um-hum" (yes) and "huh-uh" or "hum-um" (no) for anything remotely resembling these sounds of assent or denial"

Another scheme (Lander, 1996) lists several "miscellaneous words", including:

[4] This is not to say that there can be a strict ordering of activities here: on the contrary, it is not possible to fix a transcription standard without at least a tacit theory of the meanings and functions of the items being transcribed. Some thoughts on this appear elsewhere (Ward, 2000).

"nuh uh" (no), "mm hmm" (yes), "hmm mmm" (no), "mm mm" (no), "uh huh" (yes), "huh uh" (no), "uh uh" (no)

The inspiration behind these schemes seems to be the idea that grunts are just like words. This leads to two assumptions, both of which are questionable. First, there is the assumption that each grunt has some fixed meaning and some fixed functional role (filler, back-channel, etc). However, many specific grunt sounds can be found in more than one functional role, as seen in Table 1. Second, there is the assumption that the set of conversational grunts is small. However the number of observed grunts is not small, as seen in Table 2, and the set of possible grunts is probably not even finite: for example, it would not be surprising at all to hear the sound *hum-ha-han* in conversation, or *hum-ha-an*, or *hum-ha-un*, and so on, and so on. (However, not every possible sound seems likely to be a conversational grunt; for example *ziflug* would seem a surprising novelty, and would be downright weird in any of the functional positions typical for grunts.)

One concrete problem with these schemes is that they are not designed to allow phonetically accurate representations of grunts[5]. In particular, they make the task of the labeler a rather strange one. Given a grunt, first he must examine the context to determine whether it is a back-channel or a filler, then determine whether it sounds affirmative or negative, and only then can he consider what the actual sound is, and his options are limited to picking one of the labels in the functional/semantic category. The relation between the letters of the label and the phonetics of the grunt becomes somewhat arbitrary. This would be more tolerable if there was a clear tendency for each grunt to occur in only one functional position, but this is not the case, as noted above. The use of the affirmative/negative distinction as a primary classificatory feature is also also open to question. In our corpus, only 1% of the grunts were negative in meaning, and these were all in contexts where a negative answer was expected or likely, so this distinction is a strange choice for a top-level dividing principle. Moreover, negative grunts are, in fact, characterized by two-syllables with a sharp syllable boundary, often a glottal stop, and/or a sharp downstep in pitch, and/or a lack of breathiness, but these features are reflected only tenuously in the spellings listed as possible for negative grunts in these schemes.

### 2.3 Naive Transcription

The third tradition in transcribing grunts is to allow labelers to just spell them in the 'usual' way, as one might see them written in the comics or in a detective novel. The inspiration behind this is that native speakers generally have had a lot of exposure to orthographic representations of grunts, and can be trusted to do the right thing.

One problem with this tradition is that the mapping from letter sequences to the actual sounds is not clear. For example, a conversation transcription given as a textbook example of good practice includes "u" and "uh", and "oh" and "oo" (Hutchby and Wooffitt, 1999), without footnoting. Presumably the "oo" means /u/, but it could also possibly mean a version of "oh" with strong lip rounding, or a longer form of "oh", or perhaps a shorter form (if the labeler was trying to avoid confusion with the archaic vocative "o"). English orthography is phonetically ambiguous and not standardized for grunts.

A second problem with this tradition is that creaky voice (vocal fry), although pragmatically significant, is generally not represented (although many practitioners are surprisingly diligent at noting occurrences of breathiness).

### 2.4 Summary of Desiderata

Ideally we want a scheme for transcribing grunts which

1. is easy to learn and use,

[5] This is acceptable if the only aim is to train speech recognizers, where the speech recognizers' acoustic models will end up capturing the possible phonetic variation without human intervention, and if the speech recognition results are not intended for actual use, but merely to be fed into an algorithm for computing recognition scores.

| | total | back-channel | filler | dis-fluency | isolate | response | confirm-ation | final | other |
|---|---|---|---|---|---|---|---|---|---|
| [clear-throat] | 2 | . | . | 1 | . | . | . | . | 1 |
| tsk | 22 | . | 12 | 2 | 1 | . | . | . | 7 |
| ah | 7 | 1 | 3 | 3 | . | . | . | . | . |
| aum | 5 | . | 4 | 1 | . | . | . | . | . |
| hh | 3 | . | . | . | 2 | . | . | 1 | . |
| hmm | 2 | . | . | . | 1 | . | . | . | 1 |
| huh | 2 | . | 1 | . | 1 | . | . | . | . |
| m-hm | 2 | 2 | . | . | . | . | . | . | . |
| mm | 2 | 2 | . | . | . | . | . | . | . |
| mmm | 3 | 2 | 1 | . | . | . | . | . | . |
| myeah | 2 | 2 | . | . | . | . | . | . | . |
| nn-hn | 4 | 4 | . | . | . | . | . | . | . |
| oh | 20 | 6 | 9 | . | . | . | . | . | 5 |
| oh-okay | 2 | 1 | . | . | . | 1 | . | . | . |
| okay | 8 | 2 | 2 | . | . | 1 | 2 | . | 1 |
| u-uh | 4 | . | . | 2 | . | 2 | . | . | . |
| uh | 38 | . | 14 | 21 | 1 | . | . | 1 | 1 |
| uh-hn | 2 | 2 | . | . | . | . | . | . | . |
| uh-huh | 3 | 3 | . | . | . | . | . | . | . |
| uh-uh | 2 | . | 1 | 1 | . | . | . | . | . |
| uhh | 2 | . | 2 | . | . | . | . | . | . |
| ukay | 2 | 1 | 1 | . | . | . | . | . | . |
| um | 20 | . | 10 | 8 | . | . | . | 1 | 1 |
| umm | 5 | . | 5 | . | . | . | . | . | . |
| uu | 5 | 2 | 2 | . | . | . | . | . | 1 |
| uum | 5 | . | 3 | 2 | . | . | . | . | . |
| yeah | 71 | 27 | 19 | 1 | 6 | 6 | 6 | 2 | 4 |
| (other) | 72 | 34 | 19 | 3 | 8 | 3 | . | 1 | 4 |
| Total | 317 | 91 | 108 | 45 | 20 | 13 | 8 | 6 | 26 |

Table 1: Counts of Grunt Occurrences in various positions and functional roles, for all grunts occurring 2 or more times in our corpus

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| [clear-throat] | 2 | haah | 1 | nn-nnn | 1 | u-uh | 4 | unununu | 1 |
| tsk | 23 | hh | 3 | nu | 1 | u-uun | 1 | uu | 5 |
| tsk-naa | 1 | hh-aaaah | 1 | nuuuuu | 1 | uam | 1 | uum | 5 |
| tsk-neeu | 1 | hhh | 1 | nyaa-haao | 1 | uh | 38 | uumm | 1 |
| tsk-ooh | 1 | hhh-uuuh | 1 | nyeah | 1 | uh-hn | 2 | uun | 1 |
| tsk-yeah | 1 | hhn | 1 | o-w | 1 | uh-hn-uh-hn | 1 | uuuh | 1 |
| [inhale] | 1 | hmm | 2 | oa | 1 | uh-huh | 3 | uuuuuuu | 1 |
| [unsticking] | 4 | 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 | 21 | yeahh | 1 |
| aoo | 1 | mm-hm | 1 | ookay | 1 | um-hm-uh-hm | 1 | yegh | 1 |
| aum | 5 | mm-mm | 1 | oooh | 1 | umm | 4 | yeh-yeah | 1 |
| eah | 1 | mmm | 3 | ooooh | 1 | ummum | 1 | yei | 1 |
| ehh | 1 | myeah | 2 | oop-ep-oop | 1 | un-hn | 1 | yo | 1 |
| h-nmm | 1 | nn-hn | 3 | u-kay | 1 | unkay | 1 | yyeah | 1 |

Table 2: All Grunts in our Corpus, with numbers of occurrences

2. can represent all observed grunts, and

3. unambiguously represents all meaningful differences in sound.

While it is not possible to devise a single transcription scheme which is perfect for all purposes (Barry and Fourcin, 1992), it is clear that the current schemes all have room for improvement.

## 3 Proposal

The basic idea is to start with the naive transcription tradition and then tighten it up. The advantages of using this as a starting point are two. First, it's convenient, since it is ASCII, familiar, and requires no special training. Second, as the result of the cumulative result of many years of novelists' and cartoonists' efforts to represent dialog, it has presumably evolved to be fairly adequate for capturing those sounds variations which are significant to meaning.

The biggest need is to clarify and regularize the mapping from transcription to sound. This is the primary contribution of this paper: a specification of the actual phonetic values of each of the letters commonly used in transcribing conversational grunts, as follows:

**u** means schwa. This causes no confusion because high vowels, including /u/, are vanishingly rare in conversational grunts.

**n** generally means nasalization. This is unfamiliar in that English, unlike French, has no nasalized vowels in the words of the lexicon. However in grunts nasalization is common, as in *un-hn* and *nyeah*, and meaning-bearing. Occasionally there may be nasal consonants, and n can also be used for such cases, without confusion, because they appear to bear the same semantic value.

**h** generally means breathiness. This often occurs at syllable boundaries, as in *uh-huh*. Some items involve breathiness throughout a syllable, others involve a consonantal /h/, while others seem ambiguous between these two.

A single syllable-final 'h' bears no phonetic value.

**tsk** indicates an alveolar tongue click. These occur often in isolation, and occasionally grunt-initially[6].

**-** (hyphen) indicates a fairly strong syllable boundary. Phonetically this means a major dip in energy level, a sharp discontinuity in pitch, or a significant region of breathy or creaky voice.

[**repetition**] Repetition of a letter indicates length and/or multiple weakly-separated syllables.

**uu** as a syllable is a special case, indicating a creaky schwa

All other letters have the normal values.

There are two things that standard English orthography provides no way to express. These are expressed as annotations, following the basic transcription and separated from it by a comma.

**cr** indicates creaky voice, as in *yeah:cr*. For further precision numbers from 1 to 3 can be postposed, as in :cr1 for slightly creaky and :cr3 for extremely creaky.

{**numbers**} numbers after a colon indicate anchor points for the pitch contour, on the standard 1 to 5 scale. Thus *uh-uh:44-22* is a negative response or warning, but *uh-huh:43-22* is an blatantly uninterested back-channel, and *uh-huh:32-34* is the standard, polite back-channel.

Table 3 summarizes these letter-sound mappings. Table 4 suggests which sounds are most common.

## 4 Adequacy

This scheme does fairly well by the criteria of §2.4.

[6] There are cases where the click is followed by a voiced sound without any perceptible pause (with a delay from the onset of the click to the onset of voicing of 50 to 170 milliseconds).

| notation | phonetic value |
|---|---|
| *non-trivial mappings* | |
| h | a single syllable-final 'h' bears no phonetic value, elsewhere 'h' indicates /h/ or breathiness |
| n | nasalization, occasionally a nasal consonant (other than /m/) |
| tsk | alveolar tongue click |
| u | ə (schwa) |
| repetition of a letter | length and/or multiple weakly-separated syllables |
| - (hyphen) | a fairly strong boundary between syllables or words |
| *standard mappings common in grunts* | |
| m | /m/ |
| o | /o/ |
| a | /a/ |
| y | /j/, as in *yeah* and variants |
| *idiosyncratic spellings* | |
| yeah | /jeə/ |
| kay | /keɪ/, as in *okay*, *ukay*, *unkay*, *mkay* etc. |
| uu | as a syllable, indicates a short creaky or glottalized schwa |
| *annotations* | |
| :cr | creaky voice (vocal fry) |
| :1∼5 | pitch level |

Table 3: Regularized English Orthography for Conversational Grunts

| sound | number |
|---|---|
| /m/ | 56 |
| nasalization | 20 |
| /h/ and breathiness | 38 |
| clicks | 25 |
| creaky voice | 53 |
| /schwa/ | 109 |
| /o/ | 35 |
| /a/ | 5 |

Table 4: Numbers of grunts in our corpus which include the various sound components

1. As far as clarity and usability, this scheme has a direct and simple mapping from representation to the actual phonetics. It has been trivial to learn and easy to use (at least for the author; other labelers have not yet been trained).

2. As far as representational coverage, this scheme is adequate for some 97% (=306/317) of the grunts which occur in our corpus. Thus it is not truly complete, and labelers must be allowed to escape into standard lexical orthography (for things like *oop-ep-oop* and *wow*), into IPA (for cases like *achh* and *yegh*, palatal and velar fricatives, respectively), and into ad hoc notion (for cases like throat clearings and noisy exhalations).

3. As far as precision, the scheme allows sufficiently detailed representation; at least to a first approximation. In particular, it covers all known meaningful phonetic variations. It is, however possible that other phonetic distinctions are also significant. For example, it may be that the exact height of a vowel

matters, or the exact time point at which a vowel starts getting creaky, or the presence of glottal stops, lip rounding, glottalization, falsetto, and so on matter, or the precise details of pitch and energy contours matter.

Conversely, the scheme is not over-precise: all the phonetic elements represented in the scheme appear to bear meanings (Ward, 2000).

Regarding unambiguity, the scheme is an improvement but has one failing: repetition of a letter represents either extended duration or the presence of multiple syllables. As these two phonetic features are generally correlated, and the difference in meaning between them is anyway subtle, this may not be a major problem.

## 5 Open Issues

This notation assumes that the component sounds are categorical (except for creakiness and pitch), but this may in fact not be the case. Rather it may be that the phonetic components of grunts have a "gradual, rather than binary, oppositional character" (Jakobson and Waugh, 1979). This is a problem especially for nasalization and for vowels: it may be that there is an infinite number of slightly but significantly different variations. Further study is required.

Experiments with multiple independent labelers are needed to evaluate usability and measure cross-labeler agreement.

Applying this notation can be complicated by dialect and individual differences. For example, the primary filler for one speaker in our corpus was *aum*. Right now it is not known whether this is a mere pronunciation variation, perhaps dialect-related, or significantly different from *um*. More study is needed.

Other languages also have conversational grunts, for example, *ouais* and *hien* in French, *ja* and *hm* in German, and *un*, *he* and *ya* in Japanese (Ward, 1998), and it may be possible to use or adapt the present scheme for these and other languages.

## References


W. J. Barry and A. J. Fourcin. 1992. Levels of labeling. *Computer Speech and Language*, pages 1–14.

J. Hamaker, Y. Zeng, and J. Picone. 1998. Rules and guidelines for transcription and segmentation of the switchboard large vocabulary conversational speech recognition corpus, version 7.1. Technical report, Institute for Signal and Information Processing, Mississippi State University.

Ian Hutchby and Robin Wooffitt. 1999. *Conversation Analysis*. Blackwell.

Roman Jakobson and Linda Waugh. 1979. *The Sound Shape of Language*. Indiana University Press.

T. Lander. 1996. The CSLU labeling guide. Technical Report CSLU-014-96, Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology.

Fernando Poyatos. 1975. Cross-cultural study of paralinguistic "alternants" in face-to-face interaction. In Adam Kendon, Richard M. Harris, and Mary R. Key, editors, *Organization of Behavior in Face-to-Face Interaction*, pages 285–314. Mouton.

George L. Trager. 1958. Paralanguage: A first approximation. *Studies in Linguistics*, pages 1–12.

Nigel Ward. 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*, pages 464–467.

Nigel Ward. 2000. The challenge of non-lexical speech sounds. In *International Conference on Spoken Language Processing*. to appear.