Prosody in feedback
Issues on feedback possibly appropriate for human-computer interaction

Term paper

Speech Technology - (n)gslt

Karin Cavallin

Autumn 2004

Abstract

In this paper some approaches on short feedback and the prosody in such feedback utterances are surveyed. The biggest problem in e.g. implementing the prosody of feedback is that so little research is made in this area. Prosody in itself is difficult to give specific parameters since it varies between speakers and language etc. The mapping between feedback utterances and the pragmatics they convey is also scarcely investigated. This paper is an attempt to make different research groups working in different fields of human-computer interaction (ASR, dialogue systems, etc) to see that they can benefit from each other's work, and also from work made in the non-computational fields of linguistics and phonetics.

1 Introduction

This report aims at presenting the few approaches on short feedback and prosody and the pragmatics hidden behind the prosody. The term feedback is tricky in it self, since no real consensus is reached on the term, neither the term in itself nor the usage of the term. In this report the main interest is on short feedback utterances like *hm, mm, ja*, where the main purpose seem more to be to maintain the flow of the conversation than actually responding to some posed question. I choose the following view of the term, stipulating (hopefully not too controversial to anyone) it as [1]:

[1] Whether feedback is a *short answer/responsive* or an utterance which is not a turn is not discussed too much here, since it is a PhD topic in itself

- reaction to the previous utterance
- effects the flow of the conversation
- signalling grounding
- requests for clarification
- displays emotions

In this paper only feedback with lexical and prosodic conveyable aspects are considered, leaving out other feedback related information as gestures and other visual clues of bodily means as raised eyebrows and nodding head (for more on that see Cerrato (2002)).

There doesn't seem to be a consensus on what to call intonation and prosody either. The first chapter of Johns-Lewis (1986) "Intonation in discourse" does not give an understandable answer to what is what. The aim of this paper is not to answer this question, and I will use prosody as a collecting name of the features of pitch change, formant frequency, stress and durations, and specify if a special feature is intended.

2 When appropriate to produce feedback

Ward finds in the article "Using prosodic clues to decide when to use back-channel utterances" (Ward, 1996) that there are clues to when back-channelling (feedback) is appropriately produced. In Japanese "a low pitch region is a good clue that the speaker is ready for a back-channel feedback" (Ward, 1996, p.1728). It is probable that some sort of prosodic clue is present in other languages as well, not necessarily the same prosodic feature of course. Japanese is a language where a lot of feedback is used, maybe up to twice as much as in English (Maynard, 1989); this makes Japanese a very interesting language for such a study.

Ward (1996) made an experiment, where subjects were supposed to unknowingly talk to a machine. The machine produced feedback in the appropriate places, i.e. when there was a low pitch region in the subjects' utterance. The outcome of the test was that none of the subjects realized that parts of the conversation was automatically produced.[2]

[2] Wards study also supports the notion that you can listen to someone, be responsive, keeping the conversation alive without actually hearing or understand what the speaker is saying to you, since you produce feedback at the appropriate places in the "conversation". Something happening very often when e.g. trying to talk to someone who's watching TV ;o)

To implement something similar in a spoken dialogue system would most probably enhance the flow of the conversation and make people more willing to use the system on a daily basis.

3 Meaning conveyed by prosody

The little computationally approached work made on these issues at all seem to mostly been made for Japanese. Besides the study of Ward (1996) on when to produce feedback, an attempt to identify pragmatics in the prosody is made by Shimojima et al. (1998). Shimojima et al. (1998) focus on the echoic responses people tend to produce in a conversation. By echoic response is meant the phenomenon of people tending to repeat parts (or the whole) of the previous turn as a response. Shimojima et al work with the hypothesis that features in the prosody (length, timing, speed, pitch and intonation) signals degrees of understanding, and are vital to the grounding[3] of acknowledgement and repair-initiation[4].

If one can identify the prosodic features of the degree of understanding, this ought to be applicable to the echoic responses as well. Depending on the sound of the echoic response, an analysis of the grounding could to be performed[5].

This touches the idea of a word independent meaning **solely** perceived by prosody, and sometimes even overriding the lexical meaning of what is conveyed, which I find really mind-boggling.

No exact threshold where something goes from acknowledged to unintelligible was detected, probably due to the speaker dependent differences. Speaker independent aspects were found though. Shimojima et al found that "a higher pitch, faster tempo, and longer delay of an echoic response reflect a lower degree in which the speaker has integrated the repeated information, while a lower pitch, slower tempo, and shorter delay reflect a higher degree of the speaker's integration rate". This goes for Japanese, but similar methods as the ones in this study can of course be performed for other languages.

[3] *grounding* is the term for the common knowledge ground the participants in a conversation try build and maintain during a conversation.

[4] *repair-initiation* is the term used for the things people use to repair what might have gone wrong in a conversation, to improve communication efficiency and accuracy.

[5] Meaning that e.g. if hesitation and doubts can be perceived automatically, there is room for quickly repairing the eventual lack of common knowledge etcetera, enhancing the conversation flow.

3.1 What prosody to produce

Ehlich (1986) is the only one, to my knowledge, who has made an extensive survey of feedback, or *Interjektionen* and the prosodic patterns, conveying different semantic/pragmatic meanings. *Rückmeldungssignale* and *Gliederungssignale* are the most common German terms within conversational research of today. *Interjektionen* is rather the grammatical term, also covering things like exclamation words etcetera. But since Ehlich (1986) wrote his contribution as early as in 1986 the term Rückmeldung might not have been as widespread in German at that time.

Grammatik der deutschen Sprache (Zifoun et al., 1997), the German reference grammar (as authoritative for German as Svenska Akademiens Grammatik (Teleman et al., 1999) is for Swedish) has a notable coverage of the prosodic pattern of interjections. A reference grammar generally aims at describing a language for people interested in knowing and/or learning how that grammar of the specific language works. Reference grammars are historically most focussed on written language. It is noteworthy and admirable that several pages are devoted to presenting interjections. In SAG not many pages are dealing with interjections. GDS has principally completely accepted the work of Ehlich, but they provide more comprehensive schedule of the relation between prosody and pragmatics. A person wanting to learn how to speak like a German has the possibility to learn even the subtleties of feedback prosody by reading the GDS.

Ehlich (1986) has a scheme of tonal aspects added to the interjection/feedback describing the prosody with diacritic signs. They are meant to be iconically presented, and fairly self-explanatory but no durational features are provided.

Trying to really grasp the difference in sound between h́m and h̀m is neither intelligible for a reading layman, nor for a phonetician unaccustomed to the standard(?) used.[6].

The most typical interjection in German, *hm*, can carry different meanings according to Ehlich (1986)/Zifoun et al. (1997).

- ȟm ȟm' hmȟm - understanding, keep on listening
- h̀m h̀m' h́m h́m'- not-understanding, not accepting, not according to what the listener expected (Erwartungskontrast)
- h́m h̄m - securing the continuity of the conversation (Kontiuitätssicherung) to bridge realisation and planning problems
- ĥm - well-being (surprise, positive feeling, taste)

[6] Given alternatives with audible examples the difference is probably straightforward for a layman and a professional

The first two are probably the most important to detect in a dialogue system, since they convey the status of integration of what was said by the system (and vice versa) and if the second is perceived by the system, it can start with some clarifying procedures.

Hm can also work as a responsive, as an answer to a question. This is not really my working definition of a feedback word, but can be argumented as such, and is in anyway presented here for completeness.

- ȟm - positive answer
- h́mh̀m - negative answer

Reduplication plays a role as well, something which I will not discuss further, but want the reader to be aware of.

The analysis that Ehlich provides ought to be translated into other languages. The different categories of understanding, well-being, acceptance, non-acceptance etcetera can probably be re-used, and for some languages the tonal patterns as well. It is likely that the pragmatic/semantic features which the German prosody conveys is similar to at least to other Germanic languages, and out of my own experience with German, it should at least be reusable for Swedish.

4 Semantics and pragmatics of feedback

Allwood et al. (1992) distinguish three main components of communication, *speech management functions*, *interactive functions* and *focussed or main message functions*. Interactive functions are divided into sequencing, turn taking and giving and eliciting feedback where feedback is what is going to be of interest in this paper.

Linguistic feedback is defined in Allwood et al. (1992) as "linguistic mechanisms which enable the participants in spoken interaction to exchange information about *basic communicative functions*, such as contact, perception, understanding, and attitudinal reactions to the communicated content" (Allwood et al., 1992, p.1).

Allwood (1987) and Allwood et al. (1992) provides a survey, of the functions of *yes*, *no*, *m*, and *ok* in relation to the mode and polarity of the preceding utterance. The aspects of polarity can in short be explained as different feedback words, conveying correspondingly to each communicative function. Usually more than one function is the appropriate, since showing perception implies ability to hold contact etcetera. According to Allwood et al. especially the prosody of *yes* and *no* and their synonyms in spoken language convey the attitudinal reactions to a positive or negative proposition respectively. Even though Allwood et al have a very thorough go-through of feedback pragmatically, "Prosody will, however, not be treated in any detail in this paper."(Allwood et al., 1992, p.9)

The distinction between feedback and short answers is not really clear, why I choose to treat the definition of feedback according to Allwood as similar to the others in this paper, even though there seems to be a tendency to consider responsives as feedback in Allwood (1987)[7].

It is probably preferred to use Allwood's really well-considered distinctions, especially since they to a degree have been implemented in GU Dialogue Systems (Larsson, 2004) and (Larsson, 2002) [8]. The survey of Allwood is to some extent is namely comparable to the survey in Ehlich (1986) and Zifoun et al. (1997). The most ideal thing would probably be to thoroughly compare Allwood and GDS, and try to reuse as much as possible from both of them, covering both pragmatic as well as prosodic aspects.

4.1 Curled "ja"

The closest research in this field is probably the work of Lindström (1999). She has a background in socio-linguistics but has taken these prosodic aspects seriously in her conversational analysis. She recognizes that these sorts of features have been scantly examined in Swedish. Her main feature of interest for this paper, is the "curled *ja*". By this Lindström means the prosodic features of a *ja* of "a lengthened vowel and a slight rise in pitch toward the end of the syllable"(Lindström, 1999, p.140).

This *ja* can even though it is supposed to lexico-semantically be an indicator of agreeing actually mean the opposite, i.e. *no*, or rather a disagreement of the previously uttered. Lindström also surveys similar research for Finnish and English. Lindström and the ones reviewed by her seem to look at these *ja nii well* when they are attached to a longer turn. No research seems to have been made on the semantics/pragmatics of prosody of e.g. the Swedish *ja* as the only component in an utterance, when it is only question of a short feedback.

[7] This distinction is not vital for this paper anyway, but is of course of interest in general

[8] Unfortunately leaving the prosody aside, but with a promising result from a pragmatician's point of view.

5 Annotation standard

It is very difficult to describe prosody. In GDS the interjections are annotated with diacritic symbols like â, à, ă etcetera to describe the prosody. This is not completely see-through. The increasing interest in prosody has however lead to the development of more ambitious annotation standards for prosody and intonation such as ToBI. ToBI is described as "a framework for developing community-wide conventions for transcribing the intonation and prosodic structure of spoken utterances in a language variety."(ToBI, 1999). This sounds promising for the prosody interested community. If there is a good annotation standard, there ought to be possibilities to implement the prosody by somehow translating the annotations into e.g. numeric pitch references.

6 Concluding words

Doing a totally introspective test on your self you will find, that often the prosody of the utterance is more information carrying than the actual lexical word. Some sounds corresponding to what orthographically is called "hm, mm, mh" etc can be uttered without the consonants. Try to give agreeing acceptance to something, keeping your lips apart and for instance your teeth together[9]. Which consonant is helping to convey the message? None! This entails that prosody is sometimes the sole information contributor in a short utterance. The usual "hmm" is not uttered with the usual "h" but the h-sound is actually coming from the nose. Here more than prosody is working. I don't know what the aspirated nose sound is recognized as. The work of Lindström also supports the idea that sometimes prosodic semantics overrides the lexical semantics, where a "ja" 'yes' should be interpreted as a "nej" 'no'.

Since I don't have the competence to perform a thorough investigation on this topic, I just aim at pointing these things out in this paper, hoping a full-fledged speech technologist will do proper research in this area, since a lot of helpful information is lost when not considering prosody and the pragmatics it entails in these short utterances.

To improve speech recognizers and synthesizers more work on prosody needs to be done. For dialogue systems the mapping of prosody and feedback is vital to enhance the grounding and keeps the conversation flowing. If the common knowledge ground has flaws, most people

[9] You will look kind of stupid, but you will probably without any problem agree or disagree with something, using only tone, and no actual consonants or vowels.

wouldn't trust the dialogue system. To build trust between man and machine and increase usage and usability dialogue systems need improvements on the prosodic analyzer, both for naturalness and grounding.

If the system e.g. can recognize hesitation in a lexical positive answer, it might be appropriate for the system to ask some clarification questions before proceeding, or extending its turn when finding that clarification is needed.

What is obvious is that most speech technologists working with dialogue systems are generally not handling feedback as one would have wanted them with respect to pragmatics. Pragmaticians/semanticists on the other hand are not using the prosodic cues to its full potential. Since proof of discrepancy between the lexical word and the intention of the uttered (as with the curled "ja") there need to be some handling of such features. Man is not supposed to adjust to the machine.

With this paper I hope both research areas will understand the necessity of covering both the semantic/pragmatic and the prosodic aspects to get good human-computer interaction systems.

References

Jens Allwood. Om det svenska systemet för språklig återkoppling. *Svenskans beskriving 16*, 1, 1987.

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. On the semantics and pragmatics of linguistic feedback. *Gothenburg Papers in Theoretical Linguistics*, 64, 1992. University of Göteborg, Dept of Linguistics.

Loredana Cerrato. Some characteristics of feedback expressions in swedish. In *TMH.OPSR - Fonetik 2002*, volume 43, pages 101–104, Stockholm, 2002. TMH.

Konrad Ehlich. *Interjektionen*, volume 111. Linguistische Arbeiten, Tübingen: Niemeyer, 1986. Habilitationsschrift.

Catherine Johns-Lewis, editor. *Intonation in discourse.* College-Hill Press Inc., Croom Helm, 1986.

Staffan Larsson. *Issue-based Dialogue Management.* PhD thesis, Göteborg University, 2002.

Staffan Larsson, 2004. http://www.ling.gu.se/dialoglab/.

Anna Lindström. *Language as social action: grammar, prosody, and interaction in Swedish conversation.* PhD thesis, Uppsala Universitet, 1999.

Senko Kumiya Maynard. *Japanese Conversation: Self-contextualism through Structure and Interactional Management.* Ablex, Norwood, N.J, USA, 1989.

Atsushi Shimojima, Hanae Koiso, Marc Swerts, and Yasuhiro Katagiri. An informational analysis of echoic responses in dialogue. The Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, pages 951–956, Hillsdale, NJ, 1998. Erlbaum, L.

Ulf Teleman, Erik Andersson, and Staffan Hellberg. *Svenska Akademiens Grammatik.* Norstedts, Stockholm, 1999.

ToBI, 1999. http://www.ling.ohio-state.edu/~tobi/.

Nigel Ward. Using prosodic cues to decide when to produce backchannel utterances. Proceedings of ICSLP, pages 1728–1731, Philadelphia, USA, 1996. ICSLP.

Gisela Zifoun, Ludger Hoffman, and Bruno Strecker. *Grammatik der deutschen Sprache.* Walter de Gruyter, Berlin, 1997.