Synthesis of prosodic attitudinal variants in German backchannel *ja*

Thorsten Stocksmeier, Stefan Kopp

Technische Fakultät
Universität Bielefeld, Germany
{tstocksm,skopp}@techfak.uni-bielefeld.de

Dafydd Gibbon

Department of Linguistics
Universität Bielefeld, Germany
gibbon@uni-bielefeld.de

Abstract

Feedback utterances are an important part of any dialog between humans. When two or more persons talk, they use short backchannel utterances to signal understanding and interest in the conversation. Surprisingly little is known about the relationship between the accompanying prosody and the meaning of feedback perceived by the dialog partner. We present a qualitative modelling study of 12 synthesized German *ja* (yes) interjections that shows the influence of prosodic features on emotional and pragmatic perception of this kind of feedback. Listeners perceived utterances as bored, hesitant, or happy and agreeing depending on the prosodic parameters used for synthesis.

Index Terms: feedback, interjection prosody, backchannel

1. Introduction

Whenever we talk to each other, we use interjections like *hmmm*, *yes* or *uh-huh* and nod our head when listening [1]. As an element of dialog structuring, feedback is so omnipresent that it often evades a closer inspection. Backchannel utterances make up an important part of contributions in dialog given by the listener, sometimes up to 87% in certain dialog settings [2]. Feedback is a significant factor concerning naturalness of a conversation and an important factor in organizing turn-taking [3]. Talking to a human who does not give feedback is an unsettling experience, actually a dialog quickly breaks down if no feedback is given [4, p. 159].

Natural language human-computer interaction strives to improve dialog success in many ways, but a careful synthesis of feedback particles has seldom been integrated into dialog systems. As will be shown, prosody plays an important part in communicating an emotional state and a stance on the subject being talked about. Different inflections of feedback evoke very different perceived meanings of backchannel utterances. Ehlich states that since ancient times interjections have been considered an expression of *mental state* (“affectus animi” [5]). While this is a very general proposition, it partly explains why feedback is such a complex phenomenon as we shall see. If there really is a direct connection between mental states and interjections, the complexity of the former should show up in the latter.

A thorough analysis of the different elements of feedback semantics can be found in Allwood et al. [6]. They assume human communication to be a result of separable functional subsystems. For them feedback consists of methods to exchange information about four essential communicative functions in direct face-to-face communication. The functions they describe are a starting point for further analysis:

- *contact* (willingness of the listener to continue interaction)
- *perception* (willingness and capability of the listener to physically receive the message being sent)
- *understanding* (willingness and capability of the listener to understand the message)
- *attitudinal reactions* (willingness and capability of the listener to react and respond to the message including approval or disapproval).

Schulz von Thun calls such functions “Selbstoffenbarung” [7] (self-revelation) so that a message consists not only of factual information but also of information about the sender. While this is true for words, it is even more true for feedback particles. Especially verbal feedback is highly overloaded with meaning. Additionally, the backchannel system is highly timing-sensitive. A feedback utterance that is half a second off can imply something completely different than if it had occured immediately. Giving verbal feedback too early may signal impatience with the speaker. Delayed feedback suggests that the listener had to think more than usual about what has been said, possibly implying doubts. Changes in inflection may turn a harmless *yes* into a blasé reply. Repeated use of a short *uh-huh* that lacks pitch dynamics may communicate that the listener is not really interested in the topic while this feedback behavior may well be a personal trait and mean continued attention.

This paper first presents existing literature on attitudinal feedback synthesis. A flexible prosody generator for feedback is presented which allows easy creation of detailed pitch contours with the help of spline-based template shapes. Based on a small study of attitudinal judgements, the evaluation of twelve German *ja* utterances concerning emotional and conversational effect is evaluated and discussed.

2. Existing research

Synthesis of feedback utterances is a rare topic in speech research. Work on interjections in linguistics alone is scarce. For Ehlich the use of the term *interjection* itself (from Latin *interiectio,* something thrown inbetween) is a sign that the formal category already carried its alleged “linguistic illegitimacy” in its name. Studies of feedback prosody are notoriously hard to find [8] with some notable exceptions. Ehlich’s monograph on German interjections [5] was groundbreaking and its influence can not be overstated. It is cited by almost all work on German feedback and remains the single most relevant publication to date about the topic. He points out that for modern high German over 200 different interjections have been found. One of his most relevant findings is that interjections have an exceptional position between the categories of *word* and *phrase*, they “represent self-contained linguistic entities of action.” [5, p. 210] According to Ehlich, interjections share the attributes of both word *and* phrase - prosodic analysis should thus not be reduced to the aspects of one *or* the other. He puts an emphasis on the *hm* interjection which is studied in great detail. Ehlich concludes that different inflections of *hm* mainly influence the agreement/disagreement perception.

A recent study of Swedish feedback has been done by Wallers [9]. She synthesized monosyllabic backchannels (*a* and *m*) and added peaks to the pitch contour at different positions. Listeners were then asked to evaluate the synthesized interjections. While some of her results remain inconclusive, she could find significant changes in perceived meaning caused by prosody. For example an early high pitch in *a* was often evaluated as equivalent to "Oh!" which was not found for peaks at the end of the sound. Effects for a certain subgenre of interjections, so-called *affect bursts*, have also been studied recently by Scott et al. [10]. They examine recordings of human emotional expression of backchannels like *yeah* and *yuck* in isolation and conclude that all emotions except surprise and disgust involved two of three acoustic factors (envelope, pitch and spectral cues). Disgust detection depended on all three, and only surprise did *not* show up as spectral cues compared to all other emotions.

3. Studying *ja*

Because of the lack of information about how prosody can influence German feedback semantics, a listening test was set up which included 12 German *ja* (yes) feedback utterances, synthesized with MBROLA [11]. The test is a first step to explore backchannel pragmatics. The *ja* utterance is represented by two phones, [j] and [a:] (German long a). It was picked because *ja* is a very common German feedback word and because it has a strong vowel that makes the pitch curve easy to hear. The test is meant to follow the recommendation of van Bezooijen and van Heuven [12, p. 548] who encourage more concentration on the prosodic *function* than on the exact prosodic *form*. The next steps to go will be to analyze backchannels in dialog and compare results from synthesis studies like the one given here.

A spline-based pitch curve generator was created which can create smooth F_0 curves from arbitrary supporting points. The curves used were based on Ehlich's systematics of German *hm* [5, p. 304, fig. 18] and a test recording of emotional feedback inflection. Ehlich does not offer an F_0 analysis of *ja,* but for this paper it will be assumed that *ja* and *hm* prosody in German are similar enough to each other to justify adaption of the *hm* curves. The (idealized) F_0 curves used in the prosody generator are shown in figure 1. They act as template curves which can be parameterized in certain ways. For easier calculations they are normalized and lie in the 2-dimensional space of $[0, 1]^2$. The "sombrero" shape was close-copied from an emotional *ja* recorded for test purposes where it occured for happy feedback. The "hockey stick" shape appeared for even-tempered and attentive state. Dropping shape and "Ehlich's v" are taken from Ehlich's aforementioned *hm* systematics (the "Ehlich's v" curve is also found in Swedish by Wallers [9, p. 17]). The "twin peaks" curve was an attempt to generate feedback that sounded anxious. A completely linear pitch drop can be realized with the "linear fall" shape.

The parameterization concerning pitch is done by the arguments $baseF_0$ (speaker's base frequency) and $mult$ (multiplicator for the value of the pitch shape). Two parameters control timing: $duration$ (given in milliseconds) and $ratio$ (a percentage value). The latter parameter controls the length of the first phone compared to the second phone. The assumed effect of a ratio change is increased perception of hesitation by the listener. This is in line with Carlson et al. who conclude "the perception of hesitation is strongly influenced by deviations from an expected temporal pattern" [13]. Staying on the [j] phone for a longer time than usual when producing a *ja* is presumably a way to postpone communicating an evaluation in the backchannel while already giving (albeit undecisive) feedback. A curve is fully specified for the prosody generator by the tuple ($curveType$, $baseF_0$, $mult$, $duration$, $ratio$). The $mult$ value is multiplied with the current curve template value and added to the base frequency, thus controlling the curve slope:

$$curve(t) = baseF_0 + (mult \cdot curveTemplate(t)), t \in [0, 1]$$

ID	shape	$baseF_0$	$mult$	dur(ms)	j:a ratio
1	sombrero	120	60	400	1 : 5
2	sombrero	120	120	500	1 : 6,25
3	Ehlich's v	120	120	500	1 : 6,25
4	Ehlich's v	120	60	300	1 : 3,75
5	dropping	120	120	500	1 : 6,25
6	dropping	120	60	300	1 : 3,75
7	linear fall	120	25	800	1 : 6,25
8	linear fall	120	10	500	1 : 6,25
9	stick	120	40	600	1 : 2
10	stick	120	40	600	1 : 1
11	twin peaks	150	70	500	1 : 5
12	linear fall	120	25	1000	1 : 5

Table 1: Inflections of *ja* used in the listening test

The parameters used for synthesizing the test utterances are shown in table 1. A reason for choosing these parameters was to use many of the generator's degrees of freedom while keeping a manageable number of utterances. The voice used was the MBROLA `de1` voice. Subjects were placed in a silent surrounding and given the opportunity to elicit all utterances in any order and as often as they wished. The initial recommendation was to listen to all of the *ja*s to allow subjects become familiar with the choice of available utterances. Participants were asked to fill out a questionnaire with seven semantic differentials [14] for each utterance. The polar pairs used for the evaluation were: happy vs. sad, brave vs. anxious, certain vs. hesitant, approving vs. rejecting, pushing vs. not pushing, surprised vs. bored, angry vs. balanced. Each differential offered five positions to choose from, with the middle value defined as "neither one nor the other" or "not appropriate".

4. Results

A total of 12 questionnaires was submitted. All subjects were native German speakers without prior exposure to prosody listening tests. As planned, subjects found the utterances 9 and 10 hesitant, so the stick shape combined with a prolonged first phone was successful in communicating hesitation. Utterance 3 was also considered very hesitating, possibly because it has a pitch curve that is comparable to a question ("Ja?"). Yet the shape itself does not cause that evaluation alone. In utterance 4, the same shape (Ehlich's v) is used, but it is 300 ms long instead of 500 ms with a $mult$ value of 60 instead of 120. The highest pitch is thus not 210 Hz but only 160 Hz. Utterance 4 is almost unanimously evaluated as neutral in all differentials.

Certainty was especially attributed to utterances 1, 2 and

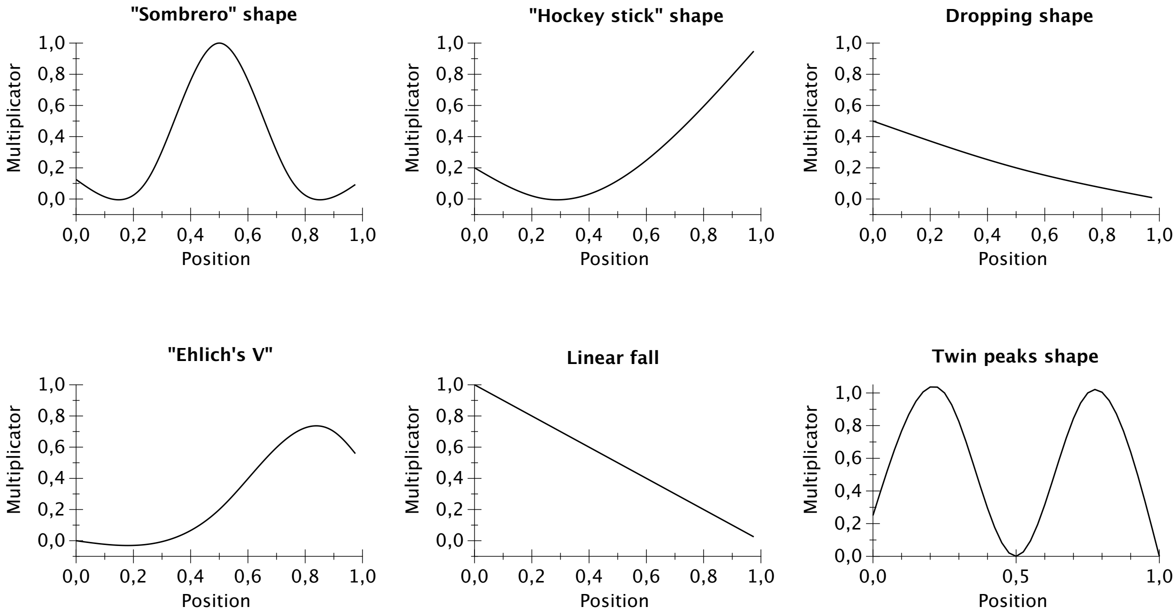

Figure 1: F_0 curves used for generating feedback

11. The first two have a sombrero pitch contour while utterance 11 has a contour with a double peak. Both have a rising-falling pitch in common (reduplicated for the double peaks).

Results for the surprised-bored differential shows interesting results. Boredness was perceived in utterances 7, 8 and 12. All three have a flat F_0 curve and are rather long (8) or very long (7, 12). Subjects were very sure about the evaluation of utterances 7 and 12 which have very flat contours (frequency span is 25 Hz). Utterance 8 has visibly more variance in subject evaluations, possibly because it has a medium duration compared to the longer utterances 7 and 12. A long duration with a flat pitch contour seems to communicate boredness.

Surprise is found in utterances 2 and 11. Characteristical for both is their high frequency span (120 Hz for utterance 2 and 70 Hz for utterance 11) as well as their rising then falling pitch contour.

The angry vs. balanced differential mostly consists of evaluations in the balanced range, but these are not very strong except for utterance 1. Why this utterance is seen as the most balanced is unclear. Utterances 7 and 12 are considered slightly angry. They have a very flat contour and a small $mult$ value in common which may implicate that a very monotonous pitch in combination with a medium or long feedback duration communicates anger. Evaluations of this differential show a high variance for utterances 2 and 5 where the angry/balance level may be difficult to judge.

Another case where results lean in one direction is the pushing vs. not pushing differential. Most utterances scored high in the “not pushing” direction while only three (1, 2 and 7) were considered neutral or slightly pushing. No utterance was considered very pushing, but variances are very high. The differential does not work too well because ”not pushing” was not a good polar attribute to “pushing”. This was confirmed by several subjects who had difficulties finding an appropriate evaluation there.

The brave-anxious differential offers interesting insights. Utterances 1, 2 and 11 point in the “brave” direction. This triple is already known from the certainty evaluations where they had high scores. Most other utterances were evaluated slightly anxious with the median at the neutral position but a relatively high variance. Only for utterances 7 and 12 were subjects sure that there is neither polar attribute in them.

In the results of the happy-sad differential only utterances 2, 11 (slightly happy) and 9, 12 (slightly sad) stand out. The reason for the impression of happiness may be the sombrero shape with a rather high frequency span which occurs in both 2 and 11.

As the German *ja* in non-feedback use means approval, results for the approving-rejecting differential are especially interesting. How far is it possible to tone down the original semantics by using prosody? A strong approval can be found for the already-known triple of utterances 1, 2 and 11. It seems like the rising-falling shape communicates this strong agreement to what has been said. The dropping shape also does this, but only about half as strong as the sombrero shape (utterances 5 and 6). Only utterance 12 is evaluated as strongly rejecting feedback. It is a long and monotone feedback that subjects instantly recognized as strong irritation (actually most subjects began to laugh instantly when hearing it because to them it sounded almost silly). Given an impression of irritation, it is understandable that subjects also evaluated the utterance as rejecting. Surprisingly, several other utterances are also evaluated slightly rejecting (3, 8, 9, 10) which clashes with the originally positive meaning of the word *ja*. For utterance 3, the pitch curve similar to a question may lead to the impression of doubtful feedback.

Subjects were invited to additionally write down their impression of certain utterances. The spectrum of answers showed that not everybody agrees on the same thing, and interpretations can become quite colorful. While utterance 4 was generally seen as very neutral (“clear and simple *ja*”), utterances 7 (“promising to not drink alcohol at a party”) and 12 (“Don’t bother me and do your job!”, “very annoyed *ja*, typical answer to motherly questions”) evoked very strong reactions. Utterance 9 was generally considered hesitating or deliberative (“unsure,

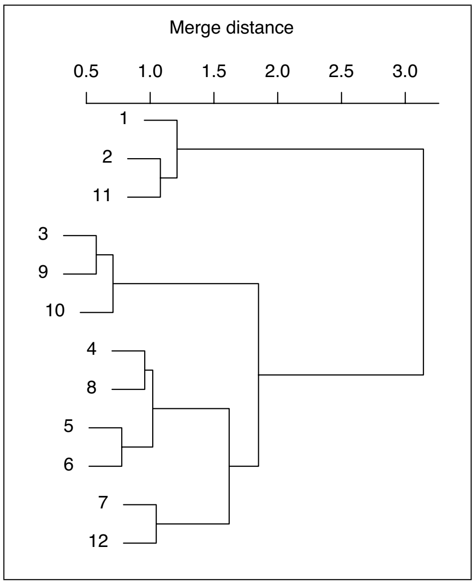

Figure 2: Clustering dendrogram

slightly anxious *ja*") Obviously the lack of context in the test invites subjects to think of possible situations where an utterance would be appropriate. The comments show that especially annoyed feedback can be associated with parent-child interaction and also the workplace (an unfriendly boss). While these evaluations remain anecdotic, they show that people associate inflections of feedback with a certain social surrounding.

Given the questionnaire results it was possible to calculate a hierarchical agglomerative clustering [15] of the *ja* utterances. Results can be seen in the tree diagram in figure 2. The Euclidean distance at which clusters were merged is displayed on the horizontal axis. The dendrogram shows that there is a cluster of utterances 1, 2 and 11 far apart from the others (the branch of the clustering tree separates at a large distance on the right). These utterances are known to clog together in the previous analysis: they are evaluated as agreeing and happy, i.e. this could be called the *agreeing and happy cluster*. A distinct cluster consists of utterances 7 and 12. The two feedbacks are those evaluated as very bored, this could be called the *boredness cluster*. Another well explainable cluster is the one formed by 3, 9 and 10. These are the utterances which were evaluated as very hesitative. One could call this the *hesitation cluster*. A detailed analysis of the data is included in [16].

5. Conclusion

As can be seen from the results of the presented study, prosody is an important factor in the perception of feedback verbals. By setting appropriate synthesis parameters, a deliberative change in the communicated semantics is possible. For *ja* hesitation could be created by prolonging the [j] phone and choosing a question-like rising pitch contour. Approving and happy feedback was found for the sombrero-shaped pitch curve. Irritation could be evoked by increasing the duration of the utterance and choosing a flat pitch shape. More research should further map the relationship between prosody and meaning in feedback. More studies on the back-channel in actual dialog are needed to further explore the topic. Attitudinal inflections of feedback can be useful in any flexible dialog system that has an inner communicative state which it can make known to a human via the backchannel. A good understanding of what feedback to produce when is useful for all systems intended to exhibit human-like conversational skills.

6. References

[1] J. Allwood and L. Cerrato, "A study of gestural feedback expressions," in *First Nordic Symposium on Multimodal Communication, Copenhagen, 23-24 September* (P. Paggio, K. J. K., and A. Jönsson, eds.), pp. 7–22, 2003.

[2] L. Cerrato, "Some characteristics of feedback expressions in Swedish," in *TMH-QPSR - Fonetik*, vol. 44, 2002.

[3] V. H. Yngve, "On getting a word in edgewise," in *Papers from the Sixth Regional Meeting of the Chicago Linguistics Society, April 16-18*, pp. 567–578, University of Chicago, Department of Linguistics, 1970.

[4] S. K. Maynard, *Japanese conversation. Self-contextualization through Structure and Interactional Management*, vol. XXXV of *Advances in Discourse Processes*. Ablex Publishing, 1989.

[5] K. Ehlich, *Interjektionen*. Max Niemeyer Verlag, 1986.

[6] J. Allwood, J. Nivre, and E. Ahlsen, "On the semantics and pragmatics of linguistic feedback," *Journal of semantics*, vol. 9(1), pp. 1–26, 1992.

[7] F. Schulz von Thun, *Miteinander reden 1 - Störungen und Klärungen*. rororo, 1998.

[8] K. Cavallin, "Prosody in feedback – issues on feedback possibly appropriate for human-computer interaction." Term paper, Nordic Graduate School of Language Technology, Autumn 2004.

[9] Å. Wallers, "Minor sounds of major importance - prosodic manipulation of synthetic backchannels in Swedish," Master's thesis, KTH Stockholm - School of Computer Science and Communication, 2006.

[10] S. Scott and D. Sauter, "Non-verbal expressions of emotion - acoustics, valence and cross cultural factors," in *Speech Prosody. 3rd International Conference Dresden* (R. Hoffmann, ed.), TUDpress Verlag der Wissenschaften, Dresden, May 2006.

[11] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. v. der Vrecken, "The MBROLA project: Towards a Set of High Quality Speech Synthesizers Free of Use for Non Commercial Purposes," in *Proceedings of ICSLP '96*, vol. 3, (Philadelphia, PA), pp. 1393–1396, October 1996.

[12] R. van Bezooijen and V. van Heuven, *Handbook of Standards and Resources for Spoken Language Systems*, ch. Assessment of synthesis systems, pp. 481–563. Mouton de Gruyter, 1997.

[13] R. Carlson, K. Gustafson, and E. Strangert, "Prosodic Cues for Hesitation," in *Proceedings from Fonetik 2006, Lund University, Sweden*, pp. 21–24, June 2006.

[14] C. E. Osgood, G. Suci, and P. Tannenbaum, *The measurement of meaning*. University of Illinois Press, 1957.

[15] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[16] T. Stocksmeier, "A multimodal Feedback Model for an Embodied Conversational Agent," Master's thesis, Universität Bielefeld, 2007.