Perception of Non-Verbal Emotional Listener Feedback

Marc Schröder
DFKI GmbH
Saarbrücken, Germany
schroed@dfki.de

Dirk Heylen
University of Twente
The Netherlands
heylen@ewi.utwente.nl

Isabella Poggi
University of Rome
Italy
poggi@uniroma3.it

Abstract

This paper reports on a listening test assessing the perception of short non-verbal emotional vocalisations emitted by a listener as feedback to the speaker. We clarify the concepts of backchannel and feedback, and investigate the use of affect bursts as a means of giving emotional feedback via the backchannel. Experiments with German and Dutch subjects confirm that the recognition of emotion from affect bursts in a dialogical context is similar to their perception in isolation. We also investigate the acceptability of affect bursts when used as listener feedback. Acceptability appears to be linked to display rules for emotion expression. While many ratings were similar between Dutch and German listeners, a number of clear differences was found, suggesting language-specific affect bursts.

1. Introduction

Human-machine interaction systems should become natural to use. They should show human-like interaction skills, including the behaviour spontaneously shown by humans when *listening* to their interaction partners. Listeners give feedback [1], signalling whether they are engaged in the conversation, whether they understand and are interested in what the speaker is saying, whether they believe the speaker, agree, and which emotions or attitudes are elicited in them by the speaker's utterance [10]. Listeners also use backchannel utterances [17] to confirm that the speaker should keep the turn.

This complex, multi-layered communication channel by listeners appears to be little studied and, if at all, only rudimentarily implemented in state-of-the art interactive systems [9].

Some research has focused on predicting the right places for giving backchannel feedback (e.g., [6], [16]). The functions of backchannels and feedback are reasonably clear; we clarify these concepts in Section 2 below.

However, little seems to be known about the *form* of listener feedback utterances. Even though the often-mentioned "mm-hmm" backchannel may be ambiguous with respect to the meaning conveyed, it seems reasonable to assume that at least some of the many functions of backchannel feedback can be linked to certain surface forms of the corresponding feedback utterances.

This paper addresses the specific issue of emotional listener feedback. As a first step to understanding this phenomenon, we simulate listener feedback by embedding non-verbal emotional vocalisations into a speaker sentence. In a listening test, we assess the emotion expressed by the feedback and the appropriateness of the feedback in the given context.

This research was funded by the EU Project HUMAINE (IST-507422).

2. Listener behaviour

When we talk to other people, we do not only want them to understand the content of what we are saying, but we also want to know how they react to it. We expect a feedback from them in order to understand the effect of our words on their mental state, their thoughts and feelings about what we are saying.

The listener may provide feedback by officially taking the floor and expressing what s/he thinks or feels in a whole turn, or may provide feedback during the turn of the speaker.

In conversation research, two notions have been proposed to deal with these aspects of communicative interaction, the notion of *backchannel* and the notion of *feedback*.

Feedback [1] may be defined as the whole set of reactions to the speaker's talk that are communicated by the addressee.

From the semantic point of view, the concept of feedback encompasses both the information about whether the listener is hearing, following, understanding what the speaker is saying and information about his/her stance towards it, including evaluations, emotions, and tendencies to act and to react.

From the point of view of its occurrence, feedback can be provided both while the speaker is holding his/her turn, and by the (former) listener taking the turn (now becoming a speaker).

Yngve [17] introduced the term "backchannel" in one of the classic texts dealing with expressive behaviours of listeners. His main concern was with turn-change signals in general. These do not only involve signals of speakers that have the intention to yield or keep the turn, but also signals by listeners indicating that they will or will not (yet) take the turn. These signals by listeners are part of the "backchannel". In this conception, backchannel behaviours, or backchannels for short, mostly function as a way to set the common ground of conversation. They do not add much new information but mainly fulfil the "control goals" [4] of conversation – the permanent goals of a talking person to know if her discourse is being attended to, heard, and understood.

Backchannels can take various forms. In some cases they use another modality, like gaze or facial expression, but short vocal expressions can also be used, such as interjections or affect bursts (see below) that do not interrupt the speaker's speech flow.

Thus, one could say that the notion of backchannel stereotypically describes a subset of feedback behaviour – those reactions provided by the listener to the speaker that, on the semantic side, inform the speaker about the listener's comprehension, and possibly agreement/disagreement, while on the occurrence side, are provided only during the speaker turn, without the listener interrupting the speaker and without taking a turn on their own.

However, the semantic and the occurrence level need not always go together, in that the "common ground – new information" opposition may not coincide with the "speaker turn – listener turn" opposition. In other words, it is not always the case that information which makes conversation "go on" necessarily takes a turn of its own. In some cases, a comment providing a substantive reaction from the listener, concerning some of his emotions or evaluations about what the speaker is saying, may come during the turn of the speaker. In this case, some listener reaction that from the occurrence point of view looks like a backchannel, is on the semantic side actually feedback in the more general sense.

3. Expression of emotions

3.1. Affect bursts: Non-verbal emotional vocalisations

Affect bursts are "very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events" ([12], p. 170). Their vocal form ranges from non-phonemic vocalisations such as laughter or a rapid intake of breath, via phonemic vocalisations such as [a] or [m] where prosody and voice quality are crucial to conveying an emotion, to quasi-verbal interjections such as English "yuck" or "yippee" for which the segmental form transports the emotional meaning independently of the prosody.

In a previous experimental study [13], we collected a range of affect bursts for each of 10 emotions, produced in isolation by German actors. On the basis of phonetic similarity, we grouped them into 24 "affect burst classes", which were classified correctly in a listening test 81% of the time on average. We also obtained characterisations of each affect burst class in terms of the emotion dimensions arousal, valence, and power. The distinction between quasi-verbal, language-specific "affect emblems" and universal "raw affect bursts", proposed by [12], was operationalised in terms of the stability of the segmental form across subjects, which we assessed in a transcription task. We thus proposed candidates for the status of "emblem" vs. "raw burst".

In this work we investigate the use of affect bursts as a way for the listener to give emotional feedback.

3.2. The role of context in emotion perception

Cauldwell [5] demonstrated that short utterances can be perceived as anger in isolation and as emotionally neutral when perceived in the context in which they were uttered. Interestingly, the perception of anger from the utterance in isolation persisted even after having heard it in context.

Similarly, Trouvain [14] showed that certain kinds of laughter are perceived as sobs in isolation, but as laughs in context.

In both cases, the difference in perception was the consequence of *extracting* a vocal expression from its original context. It is unclear whether a similar phenomenon should be expected when a vocalisation which originally was produced in isolation by an actor is inserted into a new context.

Embedding expressive vocalisations into a new context is not a straightforward thing to do, however. Inserting laughs into a speech synthesis context, we found [15] that most were perceived as inappropriate, with the exception of a very mild laugh. The details of the circumstances under which such an insertion was considered appropriate are not yet clear.

In addition, a conversational context may change the *function* of an emotional expressive display. In the case of facial expressions, for instance, [2] showed how facial displays of emotion during conversations are not the result of the emotion felt at the time of speaking, that almost all of them are symbolic parts of messages that are integrated with other communicative signals such as words, intonation and gestures. A "surprise" expression may be used, for instance, to signal disbelief. Similarly, the interpretation of affect bursts introduced into the conversational backchannel may or may not be interpreted as a comment, a symbolic act rather than the mere expression of an emotion felt. This may influence both the judgements of what is being expressed by the affect burst and the judgements on the appropriateness of the affect burst in this context.

4. Experiment

4.1. Motivation and Research Questions

The present experiment addresses the question whether affect bursts can be used by a listener to give emotional feedback to the speaker. This question has two main aspects.

1. Do affect bursts, used as listener feedback in the context of a short dialogue fragment with an emotionally inexpressive speaker utterance, convey the same emotions as in isolation?

2. How acceptable is such feedback?

One possible expectation is that affect bursts recorded in isolation are not at all acceptable for this use – that only feedback produced in an interactive setting can be acceptable. Alternatively, it could be that the acceptability can be linked to a property of the affect burst, such as its arousal – possibly, only affect bursts with a similarly low arousal as the context utterance are acceptable.

In addition to these core questions, we also tentatively investigate the role of language background.

3. To what extent do recognition and acceptability of affect bursts differ between German and Dutch listeners?

We can expect to find differences in recognition between physiology-based "raw" affect bursts and culture-specific affect emblems.

4.2. Method

For each of the ten emotion categories studied by [13], we selected two affect bursts as follows. From the 24 affect burst classes proposed by [13], we removed the four classes identified as clear cases of quasi-verbal affect emblems. The remaining 20 classes can thus be expected to be reasonably language-independent.

For each emotion, we selected the two affect bursts which were recognised best in isolation, if possible from two different affect burst classes. This was possible for all emotions except "threat" and "elation", where both affect bursts had to be selected from the same class. Table 1 lists the original recognition rates of the selected affect bursts along with their respective emotion and affect burst class.

We created the stimuli by embedding each of the 20 selected affect bursts into a neutral speaker sentence. That sentence was deliberately semantically underspecified and spoken in an inexpressive, colloquial way. The sentence was: "Ja, dann hab' ich mir gesagt, probierste's einfach mal <pause> und dann hab' ich das gemacht!" (German); "Ja, toen zei ik tegen mezelf, probeer het maar een keer <pause> en toen heb ik het gedaan!" (Dutch); "Yeah, then I told myself, why don't you try it <pause> and then I did it!" (English translation). In both the German and the Dutch sentence, the pause was 750 ms long. The affect bursts were mixed into the sentence starting at 150 ms into the pause, without modifying the pause duration. In other words, the feedback and the second part of the speaker ut-

emotion	burst	recognition (%) isol.		recognition (%) in context		acceptability	
		de	nl	de	nl	de	nl
admiration	wow	95	100	97	89	79	70
	boah	95	23	100	11	73	36
threat	hey1	95	41	70	37	26	23
	hey2	90	19	55	22	26	38
disgust	buäh	100	69	97	59	53	37
	ih	95	97	90	82	53	45
elation	ja1	85	90	90	74	51	52
	ja2	70	44	80	40	49	68
boredom	yawn	95	100	97	96	58	49
	hmm	85	81	86	85	70	51
relief	sigh	100	100	93	74	46	56
	uff	100	88	90	78	47	45
startle	int. breath	100	100	100	96	33	34
	ah	90	74	87	48	22	41
worry	oje	100	34	87	58	62	45
	oh-oh	85	71	97	65	65	45
contempt	pha	95	81	87	82	35	48
	tse	100	71	87	77	55	50
anger	growl1	90	81	80	74	37	23
	growl2	80	58	70	48	32	22
average		92	71	87	65	49	44

Table 1: Recognition results of 20 affect bursts. de = German listeners; nl = Dutch listeners. Ratings of affect bursts in isolation for German listeners taken from [13]. Acceptability ratings ranged from 0 (very bad) to 100 (very good).

terance overlapped for those affect bursts that were longer than 600 ms. All affect bursts were normalised to the same average power as the sentence into which they were embedded. In order to mask the different recording conditions between the speaker sentence and the feedback, a low-intensity white noise (at -60dB) was added to the resulting stimuli.

The test was carried out in a web-enabled setup, using the open source tool RatingTest. The 20 stimuli were presented in an automatically randomised order. For each stimulus, subjects answered two questions. In a forced choice setup comparable to the one used by [13], they identified the emotion expressed by the listener from a list of ten categories. In addition, they rated on a continuous scale the question of how well the listener's interjection fits into the dialogue.

In the German test, 30 subjects participated (15 female; mean age: 24.1 years). 11 of these took the test in a controlled setting in a quite office room; the remaining subjects took part in the test via the web. In the Dutch test, 27 subjects participated via the web (5 female; mean age: 24.2 years).

A separate group of 32 Dutch listeners also rated the affect bursts in isolation, in order to provide Dutch data comparable to the results in [13].

4.3. Results

The first observation to make in Table 1 is that the recognition rates for affect bursts in isolation are lower for Dutch listeners than for German listeners. Differences are rather small for the vast majority of bursts; only four bursts that were highly recognised by German listeners are not recognised by Dutch listeners. The two threat bursts were badly recognised, confirming the finding in [13] that the threat and anger categories cannot be fully distinguished. Also, Dutch listeners do not seem to make the clear distinction that Germans make between "boah" (expressing admiration) and "buäh" (expressing disgust), leading to a very low recognition for "boah". Similarly low is the recognition of worry "oje", suggesting that in both cases, the language-specific segmental form may be crucial to the emotional meaning.

Regarding the recognition in context, it can be seen from Table 1 that overall recognition rates are slightly lower than for perception in isolation. However, the distribution of recognition rates across categories is very similar to the perception in isolation. We conclude that the role of context on emotion recognition in this case appears to be very small.

Acceptability ratings showed clear differences between the stimuli, but the pattern is not easy to interpret. We can observe (Table 1) that ratings tend to be consistent within emotion categories. Acceptability was rated very high for admiration (leaving aside the Dutch rating of the "boah" burst not recognised as admiration), moderately high for boredom, worry, elation, and relief, moderately low for disgust and contempt, and very low for threat, anger and startle.

Interpretation is not made easier by the inherent ambiguity of the question of "good fit" that we had asked the subjects to rate. It may have been interpreted by the subjects as a general appropriateness in the context, as we had intended; one might have found it a strange reaction as a reaction to the meaning of the carrier sentence; it may also have been used to indicate technical aspects such as a mismatch between the sound quality of context and burst or the timing of the burst; or it may have been used to indicate social appropriateness in the given context, in the sense of Ekman's *display rules*: social norms prescribed by one's culture as to "who can show what emotion to whom, when." [7] The fact that the standard deviation of acceptability ratings is relatively high (29.5 on average) indicates that the rating was not an easy task.

We verified whether the observed pattern could be explained by general properties of the emotional states expressed, using linear regression tests. A regression using the three emotion dimensions arousal, valence and power as predictors accounts for only 12% of the variance. Clearly, the general properties of emotional states as captured by emotion dimensions can not explain the acceptability ratings.

Pursuing the issue of social appropriateness, we can attempt to account for the pattern found in terms of display rules. Our results can make sense if seen as a cue to display rules whose underlying logic classifies emotions both in terms of their being positive or negative and the type of goal they monitor [3, 11].

The first display rule seems to point at a general bias against expressing negative emotions. More specifically, the most sanctioned emotions are those linked to goals of aggression (anger and threat), while a somewhat lower sanction holds over negative emotions linked to goals of evaluation (disgust and contempt). Moving up to higher scores, we find worry, relief and elation, emotions linked to the goal of well-being, and then, even higher, admiration, linked to the evaluation of others. Therefore, a positive bias toward the expression of emotions may hold, first, over emotions that show a positive evaluation of the other (admiration), then positive emotions like elation and relief, and finally over negative emotions like worry. Actually, there is a common feature to elation, relief and worry

when expressed after another sentence: they may all be viewed as empathic reactions to the other's narration.

To sum up, these results might lead us to hypothesise the following display rules for affect bursts:

- display emotions that are gratifying for the speaker (admiration);
- display emotions that show empathy toward the speaker (elation, worry and relief);
- do not display emotions that show a negative evaluation of things or persons (disgust and contempt);
- do not display emotions linked to aggression (anger and threat).

Unfortunately, no clear interpretation arises for the results concerning startle and boredom. Startle could be ruled out in that it seems to be a reflex and not an emotion [8], more likely to be caused by a sudden noise than by reasoning on an interlocutor's sentence. The really puzzling result from our data is the high level of acceptability credited to boredom. This is a "cognitive" emotion, felt when the level of novel information acquired is below a minimal threshold, and does not respond to a subject's interest. It signals the low relevance of incoming information, which may be quite offending for the speaker: a cognitive emotion having quite severe social effects! On the other hand the boredom may be attributed in part to the state of the listener, such as tiredness, which makes it less offensive. So, this result is not easy to interpret in the same way as the others, which quite consistently respond to display rules coherent with rules of politeness.

5. Conclusions and Further Research

This pilot experiment investigated the effects of embedding affect bursts in a conversational setting on the judgements of their fit within this context and the recognition of the emotion conveyed by the burst compared to the recognition in isolation. We have shown that for some emotions, highly recognisable affect bursts were judged to fit well into the context.

The results of this simple test lead us to the design of new experiments, following the question of what makes a context fitting for an affect burst. We assume that at least the following issues could have an influence on the perception of emotional feedback:

a. the social acceptability of the expressed emotion, described in terms of display rules;

b. the semantic and pragmatic interaction between the speaker's utterance and the affect burst;

c. the timing of the feedback with respect to the speaker's utterance;

d. more specifically, an interaction could be expected between timing and emotion: e.g., more aroused emotions might be expressed more quickly;

e. the relation between speaker and listener and the formality of the situation.

The breadth of these aspects highlight the amount of research still needed in order to really understand emotional listener feedback.

6. References

[1] J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *Semantics*, 9(1):1–26, 1992.

[2] J. B. Bavelas and N. Chovil. Faces in dialogue. In J. Russell and J.-M. Fernandez-Dols, Eds., *The Psychology of Facial Expression*, pp. 334–346. Cambridge UP, 1997.

[3] C. Castelfranchi. Affective appraisal versus cognitive evaluation in social emotions and interactions. In A. Paiva, Ed., *Affective Interactions. Towards a New Generation of Computer Interfaces*. Springer, Berlin, 2000.

[4] C. Castelfranchi and D. Parisi. *Linguaggio Conoscenze e scopi*. Il Mulino, Bologna, 1980.

[5] R. Cauldwell. Where did the anger go? The role of context in interpreting emotion in speech. In *Proc. ISCA Workshop on Speech and Emotion*, pp. 127–131, Northern Ireland, 2000.

[6] J. Edlund, M. Heldner, and J. Gustafson. Utterance segmentation and turn-taking in spoken dialogue systems. In B. Fisseni et al., Eds., *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, pp. 576–587. Peter Lang, Frankfurt, 2005.

[7] P. Ekman. Biological and cultural contributions to body and facial movement. In J. Blacking, Ed., *The anthropology of the body*, pp. 39–84. Academic Press, London, 1977.

[8] P. Ekman, W. Friesen, and R. Simons. Is the startle reaction an emotion? *J. Pers. Soc. Psychol.*, 49(5):1416–1426, 1985.

[9] R. M. Maatman, J. Gratch, and S. Marsella. Natural behavior of a listening agent. In *5th Intl. Conf. on Interactive Virtual Agents*. Kos, Greece, 2005.

[10] C. Pélachaud et al. HUMAINE deliverable D6d: Proposal for exemplars and work towards them: Emotion in Interaction. http://emotion-research.net/deliverables, 2005.

[11] I. Poggi and M. Germani. Emotions at work. In *Proc. 8th Intl. Conf. on Human Aspects of Advanced Manufacturing: Agility and Hybrid Automation (HAAMAHA'03)*, pp. 461–468, Rome, Italy, 2003.

[12] K. Scherer. Affect bursts. In S. van Goozen, N. van de Poll, and J. Sergeant, Eds., *Emotions: Essays on emotion theory*, pp. 161–193. Lawrence Erlbaum, Hillsdale, NJ, 1994.

[13] M. Schröder. Experimental study of affect bursts. *Speech Communication*, 40(1-2):99–116, 2003.

[14] J. Trouvain. Non-verbal vocalisations – the case of laughter. Paper presented at Evolution of Language: 5th Intl. Conf., Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 31 March – 3 April 2004.

[15] J. Trouvain and M. Schröder. How (not) to add laughter to synthetic speech. In *Proc. Workshop on Affective Dialogue Systems*, pp. 229–232, Kloster Irsee, Germany, 2004.

[16] N. Ward and W. Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207, 2000.

[17] V. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–577. 1970.