Interpretation of Emotions in Natural Speech – a Comparison between Written, Auditive and Gestural Information

Åsa Abelin

Göteborg University, Sweden

E-mail: abelin@ling.gu.se

ABSTRACT

This study concerns the interpretation of emotion in natural dialogue in Swedish. The aim of this study is to investigate the role of different types of modalities. The modalities studied were: visual/written (lexico-grammatical information), auditive (adding prosodic information), and visual (adding body language). 32 listeners followed 3,5 minutes of dialogue between four persons and interpreted the emotions expressed. The results are: 1) interpretations are similar across modalities for some utterances, but vary for others. However, all listeners do not make the same interpretation in the different modalities. 2) considering all interpretations, listeners made more interpretations in the written and auditive modalites than in the visual, but the interpretations in the auditive and the visual modalities are more similar. The four most common classifications in all three modalities were very similar. 3) some interpretations were very frequent in one modality but infrequent or absent in others.

1. INTRODUCTION

What is the importance of different modalities when we interpret emotional expressions in speaker interaction? Few studies have been made of multimodal perception of emotions, and the ones that exist have often been experimental (e.g. Massaro, 2000, Massaro, 2002), while the studies done on emotions in natural conversations often are not multimodal (e.g. Cowie, Douglas-Cowie, Romano, 1999). In this study, three different modalities, or dimensions of expression, have been isolated: 1) lexical/grammatical expression (in transcription), i.e. normal multimodal communication stripped of gestural and auditive information and transformed into written form, 2) phonetic-prosodic expression, i. e. communication stripped of gestural information and 3) gestural expression in full multimodal communication. The purpose of the study is to see if interpretations of "listeners" differ or if the interpretations are the same, independently of the channels of presentation.

In many studies of emotional prosody it has been easier to find prosodic correlates to emotional speech for certain emotion words, for example for *anger*, while this has been more difficult for other emotion words (cf. Abelin and Allwood, 2000, Scherer, Banse and Wallbott, 2001). This finding can have different explanations. One possibility is that the expression of some emotions varies more than that of other emotions; another possibility is that some emotions are expressed more in a lexico-grammatical or in a gestural dimension. In the present investigation the focus was on comparing the interpretations of emotional expressions in the different dimensions. There is probably also an interaction between the dimensions, partly in the expression of emotions, partly in the interpretation. It is a well-known fact from experiments in multimodal perception that visual information interferes with auditive, cf. e.g. Massaro (2000). Isbister & Nass (1999) studied persons who interacted with computerized figures, who expressed themselves in text and body language, in an extrovert or introvert manner. They found that persons prefer to interact with figures, which are consistent in their manner of expression, e.g. extrovert both verbally and gesturally. For the expression of emotions one might then expect that the same emotions are normally expressed simultaneously in different dimensions. The results of Isbister & Nass (1999) also showed that persons preferred figures who were complementary to their own personality, for example extrovert persons preferred to interact with introvert persons. Positive emotions for a speaker, and thereby attribution of certain positive emotions or attitudes such as joy or friendliness, could consequently depend on a number of non-linguistic factors, for example, consistency in the expression of the speaker and complementarity to the personality of the listener. Attribution of emotions are then not possible to relate directly to any single part, nor to the totality of the linguistic means of expression, lexicon/grammar, prosody or body language.

There is, thus, a contextual aspect on the interpretation of expressions of emotions, which has also been pointed out by e.g. Allwood (1985) and Wichmann (2002). When listeners ascribe emotional states to the speaker there is an influence from prosody, the meaning of the utterance and the gestural communication, but also the situation, expectations, reactions from the participants in the conversation etc influence what the listener will perceive. All aspects influence the meaning of the utterance from the perspective of the listener. If this is seen as a problem which ought to be avoided in studies of prosodic correlates to emotions, the problem can however not be avoided completely by using well-controlled experimental speech, since expectations and perspective of the listener will still be there.

The fact that all aspects influence the meaning of the

ISBN 1-876346-48-5 © 2003 UAB

utterance can, on the other hand, be seen as a resource for the listener, and it indicates that to search only for invariant prosodic correlates to emotional expressions will prove futile. The expressions of some emotions are more context-dependent and the speaker also knows this, and can automatically adjust his prosody to the demands of the situation[1.] Nevertheless, it is, of course, necessary to investigate the linguistic means of expression.

Yet another study has pointed out the interaction between non-linguistic and linguistic factors in perception. Johnson, Strand, d'Imperio (1999) have shown that the sex of a visually presented face influences which vowels a listener perceives in an acoustic continuum. Even imagined sex of the speaker influences which vowels the listener perceives.

If a listener dislikes the appearance of a speaker this would then possibly influence how he interprets a certain F0-variation, or pausing, in the utterance of a speaker.
There are numerous advantages in studying natural speech and not laboratory speech and, if large variation is encountered in the interpretation of certain emotions, the explanation can be seen in contextual terms.

In experimental studies of how listeners interpret audio or video recordings, one has to remember that the listener is not a participant in the interaction and, therefore, he has access to less, or different, information than the participants of the conversation have. Consequently, he will probably make slightly different interpretations than the participants of the conversation themselves do. One can therefore imagine that some classifications which the listeners make, are more from an outside perspective of the observer, for example when someone is described as *disagreeable.* This is more of an interpersonal emotion – or attitude – while e.g. *excited* refers more to a physical state and *disappointed* refers to a more cognitive state[2]. The participants of the conversation could do other judgements of the behaviors of the co-discussants because they see them more closely, they might have body contact, a sense of smell and be more involved in the conversation.

2. METHOD

2.1 MATERIAL

The study was made on a single occasion and took approximately one hour. A video recording of 3,5 minutes discussion between four women was presented to a listener group. The recording consisted of 81 utterances[3]. The listeners were 32 students of linguistics during their first week. They were instructed to 1) study a transcription (in Modified Standard Orthography, MSO, see Larsson & Sofkova, 1996) of the conversation and to mark where and which emotions were expressed. 2) Thereafter the classified transcription was put away, another equal one was distributed and this time the test persons were instructed to listen, mark and classify in the transcription where he heard emotions. 3) The same procedure was performed a third time when the test persons watched the video recording and listened (but were told to focus on what they saw), marked and classified perceived emotions onto the transcription sheet. The whole task was open in terms of segmentation of speech and choice of emotions.

2.2 ANALYSIS

In the analysis, each utterance was seen as a unit which was given a certain classification. The classifications were registered, utterance for utterance and according to whether the classification concerned transcription only, transcription + audio or transcription + audio + video. Thereafter the interpretations in the different modalities have been compared, for the number of utterances.

Furthermore, the classification in the different modalities, in general, were analyzed in order to see if for example some expressions of emotions were attributed more in certain modalities but not in others.

3. RESULTS

3.1 ANALYSIS OF SINGLE UTTERANCES

The total number of utterances was 81, and of these approximately 10–20 utterances were classified by each listener, however not always the same utterances. For some utterances the classifications agree in general in the three dimensions, even if individual speakers vary and there are examples of classifications which are very different. For other utterances, the classifications were found to vary more in the different dimensions. Presented below are the results for one utterance, as an example, thereafter follow the results for the discussion as a whole.

Analysis of utterance 1:

"I wonder what would happen if we joined the European Union" ("ja{g} undrar va{d} som skulle hända om vi går me{d} i < eg >"). The different categorizations given are shown in Table 1 (in translation).

This utterance was interpreted and classified by 28 of the 32 listeners, in one or more of the three dimensions. Reading the transcription 21 listeners responded, listening to the tape 20 listeners responded, and watching the video, 15 listeners responded.[4]. In Table 1 it can be seen that the most common interpretations were *wondering*, followed by *questioning*, and *curious*, for all the three modalities. The interpretations may come partly from the lexical-grammatical content of the

[1] Cf. discussion in Lindblom (1983).

[2] For an overview of different theories of emotions, see Cornelius (2000), Cowie & Cornelius (2003).

[3] An utterance is defined as "a stretch of speech by a speaker, which is delimited by speech of a second speaker at the same time as the first speaker is quiet".

[4] Generally fewer classifications were made after watching the video.

word *wonder*, the question word *what,* the subjunctive of *would* and the conjunction *if*, as well as from the prosody of the utterance, and the body posture and facial expression of the talker while uttering the sentence.

transcript		transcript+ auditive		transcript+ auditive+ visual	
wondering	7	wondering	7	wondering	6
questioning	4	questioning	6	questioning	4
curious	3	curious	2	curious	1
anxiety	3	thoughtful	2	insecure	1
insecure	2	doubtful	1	observing	1
wonder	1	sceptical	1	critical	1
hesitant	1	sturdy wonder	1	expansive	1
determined	1	anxiety	1	superior	1
		positive	1		
		secure	1		

Table 1: Interpretations of utterance No 1: "I wonder what would happen if we joined the European Union"[5]

An interpretation which distinguishes the transcription from the other modalities is *anxiety*, an emotion which is more common in the interpretation of the transcript. This could be understood as prosody weakening an interpretation which builds on lexical/grammatical content, i.e. the voice does not contain any phonetic features characteristic for *anxiety* and therefore weakens this interpretation. In fact, characteristic for the interpretations in the auditive and visual modalities are interpretations which are almost contrary to the insecurity of *wondering*, namely *sceptical*, *sturdy, secure*, *critical*, *expansive* and *superior* and could be due to for example the relatively low F0 and upright body posture of the speaker.[6]

Even if the interpretations are quite similar in the different modalities this does not mean that each listener made the same interpretations in the different modalities, i.e. he does not automatically continue with his interpretation based on the transcription, but often varies as between *wondering – questioning – superior*, or *anxiety – sceptical – wondering*.

3.2 ANALYSIS OF ALL UTTERANCES

In order to see which emotions were generally attributed to the speakers in a certain modality, all the interpretations were analyzed. The number of different interpretations made were, for transcript: 31, audio: 31 and video: 19. The 10 most common interpretations are shown in Table 2.

transcript		transcript+ auditive		transcript+ auditive+ visual	
irritated	28	happy	42	insecure	13
resigned	28	questioning	37	wondering	11
insecure	25	insecure	33	questioning	8
wondering	25	wondering	24	superior	8
upset	20	negative	22	happy	7
questioning	17	resigned	21	ironic	6
happy	17	irritated	16	negative	5
negative	16	sceptical	13	self-assured	5
hesitant	15	critical	12	hesitant	5
determined	11	positive	12	irritated	4

Table 2: Interpretations of all utterances: The 10 most common classifications.

There are similarities between the modalities: The interpretations that are common in all the modalites are *insecure, wondering, questioning, happy*. Audio and video are most similar, with *insecure, wondering* and *questioning* among the four most frequent interpretations. These three interpretations are similar, connected to "not knowing", which then might be said to generally characterize the 3,5 minutes of discussion.

There are also differences between the modalities: In interpretation of transcription *irritated* and *resigned* are the most frequent, in auditive modality *happy* is the most frequent interpretation, and in interpretation of video *insecure* is the most frequent emotion. The explanation to this might be that happiness is mainly coded in the voice and not in words and grammar and thus not apparent in the transcription. Furthermore, *insecurity* might be primarily coded in gestures and body posture. However, as can be seen in Table 2, these interpretations are not absent in the other modalities, only less common. On the other hand, among the most common interpretations, there are some words that deviate more for the different modalities: *superior* is frequent for video, occurs for audio, but is absent for transcription, *irritated* is most frequent for transcription, quite frequent for audio, but quite infrequent for video. *Resigned* is most frequent for transcription, occurs for audio, and is absent for video. *Happy* is very frequent for audio, occurs for transcription and video. Thus, some interpretations are very common in one modality but more uncommon or absent in the others. A possible explanation to this is that some emotions are mainly expressed in lexical content, e.g. *irritated*, others are mainly expressed in voice, e.g. *happy*, and yet others are expressed gesturally, e.g. *superior.* A further question is if those emotions expressed in words are more

[5] Where the number of categorizations and the number of interpretations do not co-incide the reason is that some listeners have given more than one interpretation of (a part of) an utterance.

[6] Many classifications of utterances show similarities between the auditive and visual modalities and one might speculate in interdependence between prosody and body posture in production.

ISBN 1-876346-48-5 © 2003 UAB

cognitive[7], if those expressed with the voice are more physical[8], and if those expressed gesturally are more interpersonal[9].

4. CONCLUSIONS

Speakers emotions are interpreted similarly in different modalities, but there are also differences. Some interpretations are more common in a certain modality. It is possible that certain emotions utilize more certain modalities for their expression. Therefore, in the study of interpretation of emotions in communication one cannot study only the prosody of emotions, or the gestures of emotions. Different emotions may also be expressed to different degrees in the different modalities, some emotions preferring a gestural expression, others an auditive and yet others a verbal expression. There is also a need to acknowledge expression in several dimensions simultaneously, and not only expression in the auditive modality.

REFERENCES

[1] Å. Abelin and J. Allwood, "Cross linguistic interpretation of emotional prosody", *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 110–113. Newcastle, Northern Ireland, 2000.

[2] J. Allwood, "Tvärkulturell kommunikation", *Papers in Anthropological Linguistics* 12, University of Göteborg, Dept. of Linguistics, 1985.

[3] R. R. Cornelius, "Theoretical approaches to emotion", *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 3–8. Newcastle, Northern Ireland, 2000.

[4] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech", *Speech Communication*, vol. 40 (1-2), pp. 5-32, 2003.

[5] R. Cowie, E. Douglas-Cowie and A. Romano, "Changing emotional tone in dialogue and its prosodic correlates", *Proceedings of ESCA International Workshop on Dialogue and Prosody*. Veldhoven, The Netherlands, 1999.

[6] J.-M. Dubost and T. Su, "Prosodic differences and similarities between Mandarin and French in declarative, interrogative, surprise and doubt expressions", *Proceedings of the 14th International Congress of Phonetic Sciences*, pp. 1561–1564, San Francisco, 1999.

[7] K. Isbister and C. Nass, "Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics", *International Journal of Human-Computer Studies* vol. 53, pp. 251–267, 1999.

[8] K. Johnson, E. A. Strand and M. D'Imperio, "Auditory-visual integration of talker gender in vowel perception", *Journal of Phonetics* vol. 27, pp. 359–384, 1999.

[9] S. Larsson and S. Sofkova, "Modifierad Standardortografi", i S. Larsson & S. Sofkova, Eds., *Report from the HSFR project Semantics and Spoken Language*. Göteborg, 1996.

[10] B. Lindblom, "Förstå och underförstå", in *Tal och Tanke,* U. Teleman, Ed., pp. 147–178, Lund: LiberFörlag, 1983.

[11] D. W. Massaro, "Multimodal emotion perception: Analogous to speech processes", *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 114–121, Newcastle, Northern Ireland, 2000.

[12] D. W. Massaro, "Multimodal Speech Perception", in B. Granström, D. House and I. Karlsson, Eds., *Multimodality in Language and Speech Systems*, Dordrecht: Kluwer Academic Publishers, 2002.

[13] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures", *Journal of Cross-Cultural Psychology*, vol. 32 (1), pp. 76–92, 2001.

[14] A. Wichmann, "Attitudinal Intonation and the Inferential Process", *Proceedings of Speech Prosody 2002*, Aix-en-Provence, France, 2002.

[7] e.g. *irritated* is an emotional state of mind focussing on the cause of something bad.

[8] e.g. *happiness* can often be experienced physically.

[9] e.g. *superior* concerns a relation between two persons.