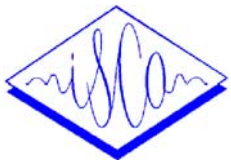

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

The Prosody of Authentic Emotions

Roland Kehrein

Forschungsinstitut "Deutscher Sprachatlas"
Philipps-University of Marburg, Germany
kehreinr@mailer.uni-marburg.de

Abstract

Prosodic variation in oral communication can occasion a shift in how a speaker's feelings and emotions are interpreted, a finding that is both commonsense and well established within the scientific community. However, neither psychologists nor researchers in linguistic and communication sciences have yet achieved clarity concerning the exact relations between the two phenomena. This paper outlines the results of a study dealing with the expression of emotions in "natural" spoken language. In this study some light could be shed on the fundamental relations between units of prosody and other signalling systems active in the expression of emotions. Results show that emotions are merely perceived as discrete and are in fact semantic composites, constructed out of several elements which each bear individual semantic components.

1. Introduction

The recognized achievements of international prosodic studies continue to stand in very poor relation to the sometimes quite taxing research efforts involved. Numerous theoretical and methodological explanations can be put forward for this. The theoretical aspects cannot be dealt with at length in this paper (but cf. [9], [3]); it is important solely to note here that, from a theoretical standpoint, a prosodic model that includes pitch alone – let alone just two relational pitch levels – is inadequate for an analysis of the expression of emotion. For this reason, I employ a differentiated, empirically founded prosodic model, which accounts for all three prosodic features (prominence and duration as well as pitch). As this model is described in detail elsewhere (cf. [3]), here I will mention just two of the descriptive categories from it, which also characterise those prosodic units found to be relevant to emotion. Functional prosodic units are classified, first, according to their domain (either *local* across syllables, or *global* across whole syntactic / pragmatic base units of spoken language; cf. [3]), and second, the form-function relation they exhibit (*discrete:* form A = meaning A, form B = function B; *continuous:* more / less form A = more / less meaning A).

Obtaining data poses a methodological problem in empirical prosody research, in particular for studies of the prosodic expression of emotions. Spontaneous, "natural" speech data collected in "natural" situations that are nonetheless suitable for acoustic analysis are needed, yet whilst the former can be obtained most readily via recordings "in the field", reliable acoustic data can best be collected in a sound studio.

In the study described in this paper, by making use of speech data that offer an optimal balance between a casual style (in the Labovian sense; cf. [4]) and acoustic quality, I have been able to demonstrate fundamental relations between prosodic units of primary linguistic relevance, prosodic units relevant to the expression of emotion, and linguistic units of verbal signalling systems in the perception of emotions in German. The empirical investigation was guided by the following research questions: 1. Why do observers / communication partners perceive certain utterances of others in context as emotional? 2. Are there prosodic units with genuine emotional meaning? If so, are they discrete or continuous units? How do these units interact formally with prosodic units that support primary linguistic functions?

2. Method

2.1. Data

The corpus upon which my research is based consists of five dialogues with a total length of 150 minutes. The conversations are all confined to the one topic, in which participants cooperatively assemble a Lego construction, one giving instructions to an unseen other. (The visual channel was deliberately excluded to help focus the analysis.) The experiment ostensibly concerned the optimisation of verbal instructions. Through various controlled modifications to the situation – manipulation of the Lego kit, the amount of time allowed, or the expectation of success – it was possible to create both a perfect distraction from the actual goal of the experiment and a situation in which participants paid minimal attention to the monitoring of speech (cf. [5]). The conversations are thus in a casual style, which Labov calls the *vernacular* (cf. [5]). Further, authentic emotions were evoked in the participants, emotions that resulted solely from the interaction and not from extraneous influences. As the speakers were seated in separate soundproof rooms and communicated with one another via microphone and headphones, they could be recorded separately on different channels of a digital audiotape. Every utterance – including overlaps – could thus be precisely acoustically analysed.

2.2. Analyses

On the basis of both the recordings and transcripts (generated according to a standardised transcription system for conversation analysis (GAT); cf. [10]), three individuals identified the emotional utterances (point of occurrence and emotional quality) in each of the five conversations. Only those utterances for which at least two of the control listeners were agreed on the classification were included in the next stages of analysis. No contradictory classifications were recorded. Conversation analytic methods were then used to describe, in context, the 374 utterances (476 instances of individual emotions) that passed the intersubjective filter, and the sources of the perceived emotions were deduced. Following what I term "the principle of holistic perception", all the signalling systems available to the interactants and observers (verbal systems and prosody) plus the context of occurrence were systematically taken into account. Hence the prosody of the emotional utterances was analysed both auditorily and acoustically (with the Praat program; cf. [2]), and the following acoustic parameters were measured: F0 mean, F0 onset, F0

offset, F0 range, F0 maximum, F0 minimum, and syllables per second. The criteria for the selection of these parameters were (1) their potential relevance in the light of previous studies (for an overview see [3]), and (2) their "translatability" into auditory qualities (relating acoustic features back to auditory qualities is a problem in many studies; cf. [1], [3]). It was decided not to calculate an average intensity since variations in this parameter can be the result of a number of factors, e.g., the distance of the subject from the microphone. Since earlier investigations indicate that deviations from a "normal" prosody are potentially relevant for the perception of emotions (yet usually base this norm on a single emotionally neutral variant of a standard utterance), an independent corpus was used here to calculate the average values for the acoustic parameters mentioned above for each of my ten subjects and provide a basis for comparison.

3. Results

Within the context of an interaction, the attribution of emotions is intersubjectively consistent. The perception of discrete emotions as semantically complex phenomena is based on a combination of all the signalling systems available during the communicative event in context. Discrete emotions are thus composites, the individual semantic elements of which are transmitted via the individual signalling systems (speech, prosody, and – in other situations – gestures, mimicry, and other body language). It proved possible to confirm the existence of emotional dimensions widely recognized in psychology as relevant semantic components. These include: *valence* (which involves the intrinsic pleasantness or unpleasantness of an event or situation; cf. [8], [11] – I distinguish between "positive" and "negative" semantic components); *activation* (this dimension concerns whether a stimulus puts an organism into a state of increased or reduced activity – here I draw a distinction between "excited" and "calm" semantic components); *dominance* (essentially, this involves whether or not individuals consider themselves able to deal with a particular situation or change and its cause – I separate the semantic components into "strong / dominant" and "weak / submissive" here); as well as the [unexpected / expected] quality (the latter records whether the situation or event that elicited the emotion was sudden and unexpected or predictable; cf. [8], [11]). Units of the signalling system of *prosody* transport individual semantic components on all of these dimensions. Empirical evidence for these units (in order of function / meaning) is offered below (for further concrete examples see [3]).

3.1. [unexpected]

Prosodically, this semantic component can be expressed through a continuous unit: a raised F0 maximum. The strength of the semantic component "unexpected" is dependent on the degree to which the F0 maximum is raised (relative to the speaker-specific reference value). The raising of the F0 maximum can be in the form of a steep rise or of an F0 jump, with a corresponding auditory shift in pitch. The following example illustrates the form of this prosodic unit (the speaker was perceived to be *shocked* and *surprised*):

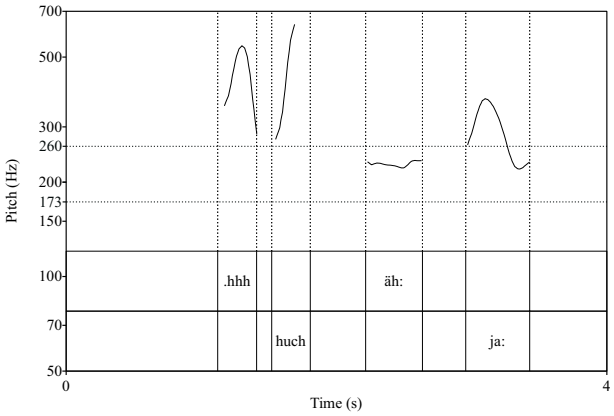

Figure 1: Acoustic analysis of utterance MA–270

The illustrated utterance, *.hhh huch,* is the speaker's (MA) reaction to her Lego bricks falling over, a completely unexpected event. This evokes what psychologists refer to as a "startle reaction" (cf. [7]). Comparing the F0 maximum at the interjection *huch* with the upper horizontal dotted line (average F0 maximum for MA) clearly shows the deviation from the reference value. In the corpus, the prosodic unit *locally raised F0 maximum* is present in utterances in which speakers are perceived as being *startled, surprised,* or *horrified.* Other emotional qualities, e.g., being *incredulous* or *delighted,* can occur simultaneously.

3.2. Activation (excited – calm)

In terms of prosody, the semantic components of emotion relevant to the activation dimension linking "excited" and "calm" are expressed through a continuous global unit: speech rate. The faster the pace of speech is, the more excited the speaker is perceived to be; the slower the pace, the more calm he or she is perceived as being. The speech rate is measured in syllables per second; pauses which might occur during the utterance are disregarded. According to van Bezooijen, this value best corresponds to the auditory impression of speech rate (cf. [1]). A clear increase in speech rate can be observed where speakers are perceived to be *excited / agitated, uncertain, eager,* or *angry*. The maximum value found in the corpus was for an utterance in which the speaker produced 9.7 syllables per second although her average speech rate was only 5.4 syllables per second. In contrast to this, conspicuous reductions in speech rate contribute to a perception of the speaker as *calm / relaxed, content,* but also as *irked* (in the sense of *demotivated*). One of the peak values for a slow speech rate in the corpus is 2.5 syllables per second (compared to a speaker-specific average of 4.6 syllables).

3.3. Dominance (strong – weak)

Emotional strength or weakness can be expressed prosodically in terms of a global unit. Implicated here is a clear and global increase or reduction in the prominence of a syntactic / pragmatic base unit. The auditory prosodic feature of *prominence* is not unproblematic in prosody studies, as its acoustic correlates have not yet been definitively established for either local (accents) or global changes in volume. According to my analyses, there is a correlation between a perceived increase / decrease in prominence and a rise / fall in the acoustic parameters of *F0 range* (relative to a speaker-specific average)

and *intensity* (relative to adjacent utterances). The following two examples illustrate, respectively, an increase and a reduction in the global prominence of an utterance.

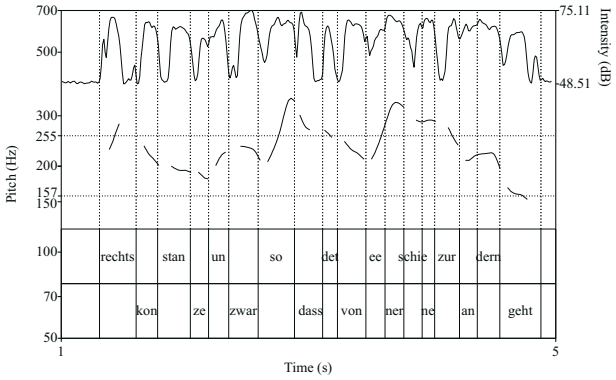

Figure 2: Acoustic analysis of utterance GE–510–512

In this utterance, the speaker was perceived as *energetic, certain,* and *angry / annoyed* (i.e., attributes which include the semantic component "strong"), which is attributable among other things to the globally enhanced prominence. Of the constitutive acoustic parameters, the increase in the F0 range is clearly apparent in the figure. Compared to this speaker's average F0 range of 97 Hz (between the horizontal dotted lines which show the F0 minimum and maximum), the range is almost twice as high (191 Hz) in the utterance depicted. Globally enhanced prominence also features when speakers are perceived as *delighted.*

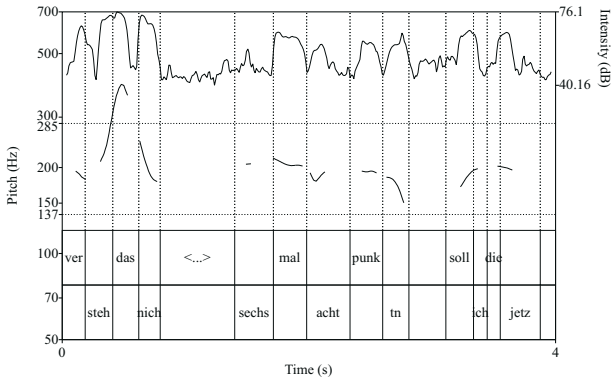

Figure 3: Acoustic analysis of utterance RU–170

In the second of the utterances shown here, the deviation in both acoustic parameters is clearly visible: both the F0 range and the intensity are markedly reduced in comparison to the reference values (the F0 range of the second utterance is 59 Hz, compared to an average value for this speaker of 149 Hz; the difference in intensity between the two utterances shown is similarly evident). The prosody of the second utterance led to the semantic component being coded as "weak" and the speaker was accordingly perceived as *uncertain* here. This prosodic unit also accompanies other perceived emotional qualities which can be assigned to the "weak" end of the *dominance* dimension (*perplexed, apologetic, resigned, frustrated, disappointed*).

3.4. Valence (positive – negative)

To date, it has been possible to demonstrate the existence of one discrete local prosodic unit for the semantic components lying on the valence dimension, with which speakers can express a positive attitude. There is currently no evidence for a prosodic unit signifying a "negative attitude / feeling". The prosodic unit that has been found is a formally complex intonation pattern whose constitutive characteristics can be explained through the following figure.

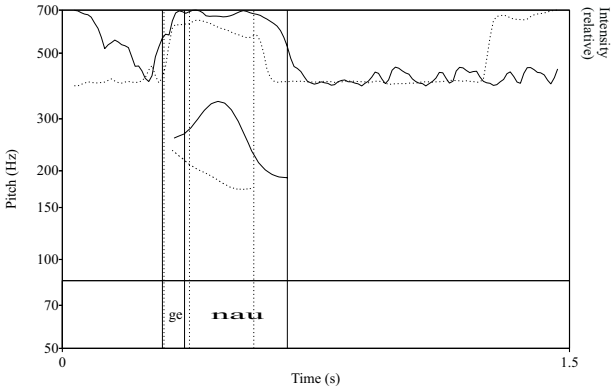

Figure 4: Acoustic analyses of utterances IS–384 (emotional utterance; solid lines) and IS–611 (unemotional utterance; dotted lines)

This figure depicts two utterances by a single speaker which are segmentally identical but prosodically distinct – to a certain extent a prosodic minimal pair. In both cases, *genau* 'exactly' is produced as an independent syntactic / pragmatic base unit with a globally falling intonation pattern. The dotted lines show the prosody of an unemotional utterance, whilst in IS–384 (solid lines) the speaker is perceived as *delighted*. It is on the second syllable of this latter utterance that the speaker produces the discrete local intonation pattern signifying a "positive attitude / feeling". The constitutive (acoustic) prosodic features of this intonation pattern are 1. temporal dilation of the syllable (perceived as stretching), 2. an F0 peak contour with a high F0 maximum (perceived as a peak in pitch), and 3. an intensity contour running counter to that of the fundamental frequency, i.e., a dip in intensity coincident with the F0 peak followed by a return to a higher intensity. This single discrete prosodic unit with its specific and complex form has also been found in other studies (such as perception experiments) and for other verbal elements. In my corpus, the local intonation pattern is associated with the meaning of "positive attitude / feeling" only when speakers are perceived as *delighted*. It is nonetheless not a "prosody of delight", since in the other studies mentioned the intonation pattern contributes to the expression of speaker attitudes such as *approval* ("Great!"), or *gastronomic pleasure* ("Delicious!"). Common to all of these uses of the intonation pattern is the fundamental meaning of "positive attitude / feeling".

4. Conclusions

In context, particular attitudes / feelings are imputed to communication partners and observed subjects with intersubjective constancy. Whilst emotions are perceived as discrete phenomena, their expression usually involves the interaction

of various signalling systems which combine to create semantic composites. Some of these individual semantic components are expressed through prosodic units. This "compositional approach to emotional meaning" (the parallel to the "compositional approach to tune meaning" of Pierrehumbert and Hirschberg (cf. [6]) is deliberate) is justified by the fact that individual prosodic units contribute to the constitution of a variety of perceived emotions. For example, an *increased speech rate* is found not just where speakers are perceived to be *excited / agitated*, but also where the emotions *uncertain* and *angry* are imputed. The differentiation between these perceived categories is dependent upon both the concurrent verbal elements and the context of the interaction. The same is also true of the prosodic units *reduced speech rate* (perceived as *calm, content,* but also as *irked* and *demotivated*) and *enhanced prominence* (perceived as *angry*, *energetic*, but also as *delighted*), and explains how, in emotional expression research to date, the same utterances could even be simultaneously classified as *angry* and *delighted* in decontextualized ratings. Methodologically, it thus becomes apparent that the assessment of emotions can only be reliably done in context.

Once the role of prosody in the communication of emotions has been adequately explained, the contribution of further vocal parameters of the speech signal – those whose specific forms are based on involuntary neurophysiological changes for example – must then be systematically investigated from this basis in future studies.

5. References

[1] Bezooijen, R.A.M.G. van, 1984. *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht / Cinnaminson: Foris.

[2] Boersma, P., n.d. *Praat: Doing Phonetics by Computer*. <http://www.fon.hum.uva.nl/praat/>.

[3] Kehrein, R., in press. *Prosodie und Emotionen*. Tübingen: Niemeyer. (Reihe Germanistische Linguistik).

[4] Labov, W., 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center of Applied Linguistics (Urban Language Series).

[5] Labov, W., 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

[6] Pierrehumbert, J.B.; Hirschberg, J. 1990. The Meaning of Intonational Contours in the Interpretation of Discourse. In: *Intentions in Communication*, Cohen, P.R. et al., eds. Cambridge, Mass.: MIT Press, 271–311.

[7] Scherer, K.R., 1986. Vocal Affect Expression: A Review and a Model for Future Research. *Psychological Bulletin,* 99(2), 143–165.

[8] Scherer, K.R.; Ceschi, G., 2000. Criteria for Emotion Recognition From Verbal and Nonverbal Expression: Studying Baggage Loss in the Airport. *Personality and Social Psychology Bulletin,* 26(3), 327–339.

[9] Schmidt, J.E., 2001. Bausteine der Intonation? *Germanistische Linguistik,* 157–158 (*Neue Wege der Intonationsforschung*, Schmidt, J.E., ed.), 9–32.

[10] Selting, M. et al., 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte,* 173, 91–122.

[11] Zentner, M.R.; Scherer, K.R., 2000. Partikuläre und integrative Ansätze. In: *Emotionspsychologie. Ein Handbuch,* Otto, J.H., et al., eds. Weinheim: Beltz, Psychologie-Verlags-Union, 151–164.