Prosody and Emotions

Sylvie Mozziconacci

Phonetics Lab.
Leiden University, The Netherlands
mozziconacci@hotmail.com

Abstract

The investigation of emotional speech constitutes quite a multidisciplinary research field. In order to benefit from the mutual enrichment that might result from this interdisciplinarity, it seems important to clarify what the specific contribution of a particular field can be. In this paper, an attempt is made to shed light on the type of contribution proposed by the tradition of studies of prosody to studies of expression of emotion in speech. Methodological issues that might contribute to the fruitfulness of studies of expressive speech are discussed.

It is argued that the way intonation conveys emotion and attitude in speech is best studied when pitch variation is represented in the theoretical framework of a model of intonation. Using such a model structures and reduces the data, formalizes the description, facilitates control of parameters and generalization of results. The use of such a model has proven useful in the field of prosody, independently of its type. Moreover, such a context constitutes an opportunity to test the model's adequacy for dealing with extreme variations such as those occurring in emotional speech. Another point is that complementary studies of production and perception are felt to be a necessary prerequisite for establishing the communicative significance of the investigated speech parameters. The need of a reference baseline in each experiment is also reminded. Furthermore, the need for drawing a distinction between the description of the global shape of the F_0 curve, i.e., the type of pith contour, and the description of how this contour is concretely implemented in terms of pitch events is also discussed. Both the abstract phonological information on the type of contour, and the concrete phonetic information on the concrete pitch implementation of the F_0 curve are pertinent. The investigation of type of pitch contour and detailed pitch implementation seems particularly promising when the experimental design is such as to allow the independent study of each effect, as well as their combined effect. This may lead us to make use of an orthogonal experimental design in the investigation of prosodic parameters. Finally, by considering how well existing assumptions about intonation account for its expressive functions, some light can be shed on the functionality of intonation for conveying meaningful information in speech communication.

1. Introduction

Speech communication conveys more than the syntactic and semantic content of sentences. Various prosodic cues are available to speaker and listener in order to encode and decode the full spoken message. In addition to fulfilling a linguistic function such as to structure discourse and dialogue, and signal focus, prosodic cues provide information such as the speaker's gender, age and physical condition, and the speaker's view, emotion, and attitude towards the topic, the dialogue partner, or the situation. Although prosodic cues other than the intonational ones, i.e. cues concerning voice quality, also contribute to the interpretation of the message, this paper will only address intonation in the investigation of speech variability conveying emotion and attitude.

A methodological difficulty in this area of research is that, at present, despite the multitude of lists of terms indicating emotions (e.g., Izard, 1977 [20]; Plutchik, 1980 [31]; Ekman, 1982 [10]; Frijda, 1986 [13]), there is no widely accepted definition and taxonomy of emotion. A review of definitions was given by Plutchik (1980, p. 81-83 [31]). Two main tendencies in theories have developed. One tendency is to consider emotions as discrete categories (Izard, 1977 [20]; Plutchik, 1980 [31]; Ekman, 1982 [10]). A distinction is made between basic emotions and combinations of these basic ones. Another tendency is to view emotions as characterized by progressive, smooth transitions (Schlosberg, 1954 [39]). Similarities and dissimilarities between emotions are characterized in terms of gradual distances on dimensions such as pleasant/unpleasant, novel/old, consistent/discrepant, control/no control. In the present issue, Zei (2002 [43]) proposes a unified theory of emotion and cognition integrating emotion, cognition and behavior into a single model of the adaptive functioning of the human organism. This model describes a three-dimensional affective space involving the dimensions: valence, arousal and power.

Presently, in studies of emotional speech, the term emotion is used with different meanings, and sometimes includes notions such as attitude, feeling, or intention. As a consequence, a large variety of emotion labels is used in these studies. Moreover, as to speech material, the lack of agreement on a definition of emotion also reflects on collection of databases. Procedures for recording of genuine spontaneous emotional speech, elicited emotional speech, and enacted emotion have to be different. The recording of genuine emotion raises ethic issues, as well as difficulties with emotion labeling, and control of the recording situation. A widely accepted recording protocol lacks also for the recording of elicited emotions, but there are guidelines for the induction of emotion in speakers, and various ways of eliciting emotions have been used in literature. However, despite the fact that emotions are not well defined, progress is made in research focusing on speech conveying emotion, mostly thanks to the ability of speakers and listeners to rely on empirically based notions of what

emotions are. This follows from the fact that subjects are quite capable of performing the pertinent experimental tasks in a consistent way. Moreover, it should be appreciated that a single emotion can be uttered in different ways.

Motivations for conducting research in the area are both theoretical and practical. The permanent evolution of expectations in the field of speech technology have led to an increased awareness that we will have to cope with speech variability and evolve towards naturally sounding speech. Indeed, nowadays, expectations include dialogue systems allowing adequate human-machine interactions. This implies a certain adjustment of the style of the human-system interaction to the behavior of the human user (de Rosis and Grasso, 2000 [34]). Unfortunately, the lack of proper prosody in synthetic speech seems to lead to uninvolved and unnaturally sounding speech. It is commonly presumed that adding variation to synthetic speech would enhance its naturalness and its acceptability. Moreover, the synthesis of emotionally colored speech can be considered as a goal, but at the same time synthesis can be used as a research tool (e.g., Carlson, 1991 [6]; Beckman, 1997 [3]). Finally, in order to facilitate the mutual enrichment of interdisciplinarity, it seems important to clarify what is the specific contribution of a particular field. In this paper, an attempt is made to shed light on the type of contribution proposed by the tradition of intonation study to the study of the expression of emotion in speech. This contribution is presented as a discussion of methodological issues that might contribute to the fruitfulness of investigations of expressive speech.

The importance of intonation as a medium for conveying emotion in speech was stated in numerous studies (e.g., Williams and Stevens, 1972 [42]; Cosmides, 1983 [8]; Cahn, 1990 [5]; Kitahara and Tohkura, 1990 [21]). Generally speaking, research concerned with emotional speech investigates speech variability. In this context, it seems logical, in such studies, to investigate the intonation variations against the background of intonation phenomena known and described in established theories. Variations conveying emotion are part of the whole set of intonation variations that are relevant to human communication, and therefore need to be dealt with by these theories. In this light, paying attention to methodological issues appears relevant for the study of intonation in emotional speech. In this paper, a few claims are formulated, discussed, and illustrated by means of examples of approaches to intonation conveying emotion.

2. Intonation conveying emotion in speech

2.1. Using a reference baseline

The need for a baseline in every experiment is certainly not controversial. However, the context of studies concerned with speech conveying emotion is interdisciplinary. The research topic constitutes a crossroads of diverse disciplines such as linguistics, psychology, neuro-psychology, ethnology, cognition, artificial intelligence, information technology, acoustics, physiology, and phonetics. In such a multidisciplinary research field implying the coexistence of various traditions of research, this crucial point of a baseline does not always appear trivial. Moreover, a baseline can certainly not be missed in such studies addressing speech variability.

Different types of references appear useful in studies of emotional speech. Naturally, in order to insure unequivocal experimental results, a reference condition is needed in each experiment. In the specific context of investigations of emotional speech, a reference baseline of non-emotional speech, most frequently called 'neutral speech', has proven useful. It allows for comparisons, not only within studies, but also across studies. Consequently, the lack of such a reference category in some studies (e.g., Banse and Scherer, 1996 [2], Protopapas and Lieberman, 1997 [32]) may allow alternative explanations, thereby compromising the unequivocal formulation of results. The lack of reference also precludes quantitative modeling of the acoustic properties relative to the reference, which renders comparisons with other studies rather difficult.

Additionally, comparisons between various speech material such as between natural speech and synthetic speech, or between spontaneous emotional speech and elicited emotional speech appear quite useful. Williams and Stevens (1972 [42]) clearly illustrate this point. In this study involving recordings from real-life situations, speech recorded in a neutral situation was used as a baseline. Moreover, and even more remarkable in this early study of spontaneous emotional speech, spontaneous emotional speech was used as a baseline for simulated emotional speech that is easier to obtain in a controlled situation.

2.2. Investigating emotional speech in the framework of a model of intonation

An important claim, as to how to carry out investigations of emotion and attitude in speech in the most fruitful way, is that such studies benefit from the theoretical framework of a model of intonation. Each model, independently of its type, constitutes a tool for representing and interpreting relevant data. A model formalizes the description, structures and reduces the data. This makes it possible to exploit synthesis as an investigation tool. Systematic manipulations can be carried out on the speech material, which is beneficial to the study of emotional speech. Results can be interpreted in terms of a model of intonation, parameters can be controlled in those terms, and procedures can be systematic within this framework. The use of such an approach involving the use of an intonation model, and systematic manipulations of parameters by means of synthesis, results in the reduction of possible alternative interpretations. Additionally, the combined use of analysis and synthesis allows for establishing a correspondence between production and perception. Moreover, studies considering speech variations as extreme as the ones occurring in the expression of emotion, are sources of opportunities for confronting measurement procedures and models commonly used in prosodic studies, with speech samples displaying a wide range of variations. If a model is found to be adequate for the description of the variations perceptually relevant to the expression of emotion in speech and for the re-synthesis of the emotional speech, its adequacy can be confirmed. On the other hand, if the model appears either to be insufficient for describing speech variations or for re-synthesizing emotional speech, then considering in which

way the model should be modified can contribute to our understanding of speech variation.

The investigation by Mozziconacci (1998 [28], 2001 [29]), aiming at seeking optimal values of prosodic parameters for conveying emotion in speech, illustrates this claim concerning the usefulness of a theoretical framework, and checks the model's validity. This study investigates intonation in production and perception of Dutch speech conveying six emotions or attitudes: joy, boredom, anger, sadness, fear, and indignation, against the 'emotion category' neutrality as a reference. The analysis of natural speech, the manipulations of natural speech by means of analysis-resynthesis, and the speech synthesis were carried out within the framework of the IPO model of intonation ('t Hart et al., 1990 [17]). This allowed for controlling parameters, enhancing the systematic aspect of procedures, and testing the adequacy of the model for processing emotional speech. The database used in this study consists of 315 utterances (3 speakers × 5 sentences × 7 elicited emotions × 3 times). On basis of the perceptual identifiability of the emotions, these utterances were selected as representative speech material. Values were experimentally derived for the generation of emotional speech from a neutral utterance, and perceptually tested in re-synthesized speech and in synthetic speech. The percentages correct identification of emotion are reported in Table 1 for three types of stimuli, used in three different experiments: 1. (close-copy stylization of) natural speech, 2. speech generated by manipulating pitch level and pitch range of natural utterances, and 3. rule-based synthetic speech involving rules for pitch level, pitch range, and speech rate. Considering that in this type of study, a typical percentage of identification of emotion in natural speech is approximately five times higher than chance (Siegwart and Scherer, 1995 [35]), and that chance level is 14.3 in this particular study using a seven-alternative forced choice paradigm, the results can be considered acceptable. Hence, this procedure does not simply provide parameter values that can be considered perceptually optimal, but it shows that the model used in the study can be considered adequate for the purpose of the investigation: the analysis and synthesis of a broad range of expressive speech.

Table 1: Percentage correct identification of emotion in Mozziconacci (1998 [28], 2001 [29])

Results of three experiments: 1. using natural speech of a male speaker, 2. using speech resynthesized after manipulation of pitch level and pitch range, 3. Using speech generated by rule-based diphone synthesis.

	Natural speech	Resynthesized speech	Synthetic speech
Neutrality	85	67	83
Joy	62	72	62
Boredom	92	85	94
Anger	32	42	51
Sadness	97	75	47
Fear	60	42	41
Indignation	85	77	68
Mean	73	66	63

Moreover, it is speculated that other intonation models would also be adequate. Higuchi et al. (1997 [19]) carried out an experiment seeking optimal values for pitch level, pitch range, and speech rate for the four speaking styles: normal, hurried, angry, and kind. Thirty-five utterances were considered in each speaking style. This time, the framework of Fujisaki's model of intonation (1991 [14]) was used for the analysis as well as for the synthesis of speech. The results, yielding high percentage identification of the speaking styles confirm the adequacy of this model for describing and generating speech deviating from the expression of neutrality.

Another point is that the framework of an intonation model influences the way notions such as pitch level and pitch range are estimated. Most frequently, such estimations are strictly data-oriented. Measures of mean F_0 and standard deviation of F_0 are common use. However, they must be expected to obscure a substantial part of the variation present in the speech material. Indeed, they do not provide information on the course of pitch within utterances, and they do not provide any direct information concerning the linguistically relevant variation. Obviously, alternative measures have also been used, and evaluated. Patterson and Ladd (1999 [30]) have considered various measurements for estimating pitch level and pitch range in speech conveying affect. A characterization of these notions was based on mean values of specific linguistic targets, and features were described as relative to the speaker's range. F_0 values were extracted at locations corresponding to sentence initial peaks (H), other accent peaks (M), valleys (L) and sentence final low (F). Two possible measures: L and F were considered for the notion 'bottom of the range', which can represent pitch level. Four options were considered for pitch range: H-F, H-L, M-F, and M-L. The correlation was investigated between these measures and the normalized perceptual judgment data concerning the categories: confident, tense, harsh, expressive, deep, weak, irritated, happy, afraid, relaxed, emphatic, and bored. The investigation involved 2 passages of text read aloud by 32 speakers, and 48 listeners participated in the listening test. The results show that a linguistically motivated approach better captures variation in listener's judgments than a statistical approach. The utterance final-low was found to be the best choice for measuring level, and the M-L measure the best for measuring range.

Mozziconacci (1995 [27], 1998 [28]) interpreted production data on pitch level and pitch range in the context of two different models. The same database was used as in Mozziconacci (1998 [28], 2001 [29]). The data were first represented as F_0 targets at anchor points, using a tonal approach, and then described within the IPO model of intonation. The differences that were observed in data representation concern two factors: relative peak heights, and final low. The pitch-accent peaks realized by speakers in the course of an utterance were not of equal excursion size, while the IPO model represents a constant excursion size of the pitch movements over the whole utterance. The final pitch value measured in the natural utterances was frequently not aligned with the end of an estimated baseline such as the intonation baseline used in the IPO model. Should these two factors be proven to be relevant to conveying emotion in speech, it would indicate that the independent modeling of these factors using tonal targets is an attractive option. An identification experiment was then carried out in order to test the perceptual

relevance of final lowering and relative height of pitch-accent peaks. Differences between results obtained in the experimental conditions, and averaged over all emotions did not reach significance. Therefore, the possibility that that these details are not very important for conveying most emotions in speech could not be rejected. This suggests that the IPO model, by providing a simplification of the pitch phenomena on the basis of perceptual relevance, simply performs a data reduction. In that case, checking the model validity did not lead to conclusive results.

Summarizing, the results of the two previously discussed studies (Patterson and Ladd, 1999 [30]; Mozziconacci, 1995 [27], 1998 [28]) support the idea that an approach based on theoretically grounded estimations is preferable to one based on data-oriented statistically grounded estimations.

2.3. Supplementing production and perception studies

Considering the need, in investigations concerned with emotional speech, to establish the communicative significance of the relevant speech parameters, it seems quite appropriate to supplement studies of production with studies of perception. To start with, the complementarity of the production and the perception processes is the basis of the spoken communication process itself. Additionally, the two types of investigation provide supplementary information. Some variations can be observed in the production data. It does not necessarily mean that they are relevant to the perception of emotion. Micro-prosody, for instance, is a typical example of variations lacking melodic perceptual relevance. Moreover, the perceptual relevance of a cue observed in a production study may be obscured in a perception study by the experimental set-up itself, or by the effect of another cue. Indeed, the efficient use of a particular parameter can allow the relaxation of another one. The convincing use of a specific parameter can cause a less saliently used parameter to remain underestimated or even unnoticed. That other parameter can be potentially relevant, though. The risk that a relevant parameter remains unnoticed is particularly present if this parameter have a small statistical effect, even if a small effect may be very pertinent. Finally, the perceptual relevance of a cue observed in production data can simply stay unnoticed because the cue can be relevant to an aspect of the communication process other than the one under study. Summarizing, it seems useful to remind that the results of production and perception studies are supplementary, and that the correspondence of results obtained in both types of study firmly establishes the communicative importance of the parameters being studied.

This need for a combination of both production and perception studies, in order to make progress in understanding vocal communication of emotion and/or attitude, was already expressed, for instance, by Scherer (1991 [37], 1996 [38]). Indeed, he argues in favor of the use of the Brunswik's lens model. This psychological framework for analyzing and modeling judgment tasks involves encoding as well as decoding strategies.

Unfortunately, most studies concerned with emotional speech focused either on production (Fairbanks and Pronovost, 1939 [11]; Williams and Stevens, 1972 [42]; Kitahara and Tohkura, 1992 [22], Amir, Ron, and Laor, 2000 [1]) or on perception (Lieberman and Michaels, 1962 [26]; Ladd, Silverman, Tolkmitt, Bergman, and Scherer, 1985 [24]; Cahn, 1990 [5]; Carlson, Granström, and Nord, 1992 [7]; Schröder, 2000 [40]) of emotion in speech. This is regrettable as it seriously undermines the generalizability of the results. As a consequence, e.g., the interesting results of Cahn's work (1990 [5]) are hard to transpose to another synthesizer because they do not relate to production data. On the other hand, in a study such as Amir et al. (2000 [1]), the results of the production study, as detailed as they can be, were not shown to be of any perceptual significance.

Summarizing, the value of the acoustic data resulting from the analysis-oriented studies, appears to depend strongly on what is known about the correspondence between the production and the perception of emotion. Establishing a relationship between acoustic and perceptual data allows for a systematic approach. Regularities and systematic differences in prosodic parameters can be extracted from the production data, hypotheses can be extracted, and the question whether it is possible to distinguish each emotion from others on this basis can be investigated. A perception study makes it possible to check the perceptual relevance of these findings. Therefore, an approach involving successively the analysis of natural speech, the re-synthesis of speech after systematic manipulation of a prosodic parameter in natural speech, and the generation of rule-based synthetic speech, seems to constitute a valuable methodological background for the study of emotional speech.

2.4. Distinguishing phonetics from phonology

The claim formulated in this section is that optimally, both the abstract information on the type of contour, and the information on its concrete pitch implementation in terms of pitch level and pitch range should be distinguished, and considered in parallel. In other words, the phonological choice of contour type, and the concrete phonetic realization of the contour both have to be addressed. The proposition formulated here is that it would be best to consider them independently of each other, as well as in combination with each other. As a consequence, the use of an orthogonal experimental design seems particularly recommended.

Intuitively, the claim formulated above corresponds to our experience with stress-accent languages, in which speakers can exploit various pitch contours in order to lend prominence and signal a boundary. Different fine meanings that are reflecting variations in emotion, attitude, and intention can be conveyed, not only with different contour types, but also with a single contour type yielding constant characteristics as to prominence-lending and boundary-signaling. Therefore, it seems sensible to assume that both the type of pitch contour, and the specific implementation of the F_0 curve in terms of pitch level and pitch range are used by speakers and listeners for the communication of the whole spoken message.

This claim is discussed here for stress-accent languages. In such languages, various pitch movements can be used for realizing a pitch accent. However, the number of pitch movements available, and the characteristics of these movements vary from a language to another. Therefore, questions will rise in the following, concerning the linguistic and para-linguistic value of the use of specific types of pitch contour in speech conveying emotion, attitude, or particular

intentions. Stress-accent languages, as they have actually been extensively investigated, seem a convenient choice for attempting a distinction between linguistic and para-linguistic matters. Naturally, characteristics of the pitch movements vary from one type of language to another. Obviously, the claim is not interesting in the case of pitch-accent languages such as Japanese. Indeed, in such languages there seems to be no choice in contour type; one single contour type is used for the realization of the accent. Nevertheless, and although this assumption was not tested, it seems reasonable to anticipate that the claim would hold for other types of languages, such as delimitative accent-languages, e.g. French or Indonesian.

Mozziconacci (1998 [28]) carried out two different types of analysis that can constitute an illustrative example of the proposal to consider contour type and phonetic realization of the F_0 curves orthogonally. The first type of analysis, concerned with the concrete realization of contours, was already briefly discussed in section 2.2. For the second analysis, focusing on the shape of the F_0 curves, a labeling was carried out using the intonation grammar for Dutch by 't Hart et al. (1990 [17]). The distribution of the pitch contours in the initial and the final parts of utterances was investigated over the various emotions. The types of contour appeared not to be equally distributed over the emotions. In order to test the effect of contour type on identification of emotion, two perception experiments were run. In a first one, pitch values were varied systematically, while the contour type was kept constant across emotions. All stimuli were generated with the '1&A 1&A' contour, i.e., two prominence-lending rise-falls as labeled in the IPO grammar of intonation for Dutch. This contour may also be transcribed as %L H*L H*L L% in the Transcription of Dutch Intonation (ToDI), the ToBI-like system developed for Dutch (Gussenhoven, forthcoming [16]). All emotions could be reasonably identified in the speech material, and well above chance. A second perception experiment, involving four conditions, was run in order to test the combined effect of the contour type, and the exact phonetic implementation of the F_0 curve. Per emotion, conditions 1 and 2 test the effect of optimal 'pitch implementation', averaged over all types of pitch contour, and averaged only over the '1&A 1&A' contour, respectively. Condition 3 tests the effect of contour type, and Condition 4 tests the combined effect of contour type and 'pitch implementation'. In comparison with each other, percentages of correct identification of emotion (See Table 2) reflect the significant main effects of pitch and contour type on subjects' responses. Percentages are low because stimuli were not prepared as instances of any intended emotion, but rather as instances of all possible combinations of phonological contour type and phonetic implementation.

Table 2: Percentage correct identification of emotion per condition in Mozziconacci (1998 [28])

Cond. 1: all contours combined with optimal pitch (per emotion).
Cond. 2: only the '1&A 1&A' contour combined with optimal pitch.
Cond. 3: optimal contour (per emotion), combined with neutral pitch.
Cond. 4: optimal contour, in combination with optimal pitch.

	Cond. 1	Cond. 2	Cond. 3	Cond. 4
Neutrality	37	46	56	56
Joy	18	10	27	35
Boredom	41	48	19	52
Anger	9	10	27	23
Sadness	8	0	33	19
Fear	19	17	10	25
Indignation	24	19	4	52
Mean	22	21	25	37

The detailed data concerning all subjects' responses were subjected to a three-way log-linear analysis (Fienberg, 1980 [12]). The data could be fitted into a log-linear model, in which significant interactions were found between 'pitch implementation' and subjects' response, and between contour type and subjects' response, but not between 'pitch implementation' and contour type ($\chi^2_{420} = 459.7$, $p > .9$). Moreover, considering the mean percentages, and assuming a chance level of 14% corresponding to 7 emotion categories, the contribution of optimal 'pitch implementation' is limited to +8%, while the contribution of phonological choice of contour type is greater, i.e., +11%. The effects of phonological choice of pitch contour and phonetic optimization of pitch seem to be roughly additive: 14% (chance), 11% (contour type), and 8% (pitch) add to 33%, which closely approximates the percentage correct identification obtained in condition 4. This, again, reflects the independent effects of phonetic 'pitch implementation', and phonological contour type on the identification of emotion. Results show that the phonological choice of contour is of primary importance for determining the emotion perceived, and that it is at least as important as the phonetic implementation in terms of pitch level and range, which relevance for conveying emotion in speech is well established.

Finally, in this investigation, the F_0 curves could be described with the IPO approach, yielding its adequacy for describing F_0 curves conveying emotion. An extension of the intonation grammar for Dutch was not necessary, and the distinctive features used in this grammar appeared to be sufficient.

One may wonder whether making use of another approach to intonation would have led to results similar to those obtained with the IPO approach. A simple comparison between intonation labeling in the IPO approach and in the auto-segmental phonological approach is not obvious, but is feasible. It is, however, beyond the scope of this paper. Nevertheless, it seems reasonable to expect that even if an auto-segmental transcription system makes different distinctions than one based on an experimental-phonetic approach, it would most probably lead to very similar general conclusions. Moreover, using a different approach to intonation does not seem to preclude a separate description of the abstract pitch contour and its concrete phonetic

implementation. An example of separate descriptions of phonetic realization and contour type in emotional speech is given by Roach (2000 [33]). In that case, the ToBI system is used for labeling a database.

The pertinence of the intonation contour in conveying meaning was also demonstrated in Grabe et al. (1997 [15]). In this study, an orthogonal design was used, combining high and low preheads with three Dutch pitch accents: H*L, H*LH, and L*H, which resulted in six experimental intonation contours. A perceptual experiment was conducted in order to test which of these contours would best convey friendliness, aloofness, self-evidentness, irritation, uncertainty and politeness. It was shown that choice of prehead conveys meaning, and a significant interaction was found between prehead and first pitch accent.

Scherer et al. (1984 [36]), by testing assumptions concerning intonation theories, reached conclusions very similar to Mozziconacci's (1998 [28]). Scherer et al. (1984 [36]) describe two approaches, qualified as the 'covariance' and the 'configuration' approaches, and test them experimentally. The approaches correspond to underlying theoretical assumptions concerning intonation. The covariance approach reflects the assumption that information on emotion function independently of information on the strictly linguistic content of utterances. According to this view, the treatment of linguistic and paralinguistic matters could be done in parallel. The first question put to the test was whether listeners' judgments of emotion in speech are based on the covariance of continuous, scalar variables with the emotion expressed and the intensity of the emotional state of the speaker. As for the configuration approach, a first underlying assumption is that the type of pitch contour is a linguistic element. Indeed, this latter view distinguishes between linguistic and paralinguistic function of F_0 variations. Another assumption is that the intonational cues conveying emotion in speech partly depend on the combination of sentence type and type of pitch contour. In other words, the type of pitch contour used in an utterance would only provide information concerning the emotion of the speaker if processed in interaction with grammatical features of the spoken text. Whether listener's judgments are based on configurations of categorical variables was put to a test in order to check the validity of the second approach.

Two experiments were conducted with speech materials composed of questions conveying politeness, impatience, reproach, hesitation, friendliness, relaxation, understanding, doubt, and aggressiveness. In the first experiment, subjects' judgments of recorded utterances were compared with judgments of written scripts. Subjects were asked to indicate, using a rating scale, which emotion was conveyed. In the second experiment, the perception test was carried out using stimuli rendered unintelligible by means of either low-pass filtering, random splicing, or reversed order. Results were compared with each other and with identification in full-audio utterances. Pitch contours were classified as rise or fall, and questions were classified as wh-question or yes/no questions. Finally, results were subjected to multiple regression analyses. Scherer et al. (1984 [36]) concluded that both the covariance and the configuration approaches should be included in any adequate general account of intonation. The relevance of overall F_0 and range was demonstrated, which argues in favor of the covariance approach. However, the influence of contour type on the perception of emotion appeared to depend on sentence type. Hence, it was also concluded that a distinction should be brought between linguistic and paralinguistic features of F_0. Therefore, results were not conclusive in terms of refuting one of the two approaches, but it was suggested that the covariance approach would be more adequate for considering speech affected by biological factors, as it is the case in physiologically based emotional states, while the configuration approach would be more appropriate for speech affected by socio-cultural and linguistic conventions. This last point will be discussed in the following section.

2.5. Differentiating between linguistic and paralinguistic value of intonation variations

Prosodic variations fulfill various functions in the communication. Some of these functions are traditionally considered linguistic, and others paralinguistic in nature (e.g., Ladd et al., 1985 [24]; Di Cristo, 2000 [9]). Although the prosodic function of conveying the expression of emotion seems to involve both a linguistic and a paralinguistic component (e.g. Laver, 1995 [25]), it is frequently considered a paralinguistic function, despite the doubts expressed on the subject (van Heuven, 1994 [18]).

Distinguishing the contour type from its detailed implementation in terms of pitch level and pitch range may well lead to a distinction between linguistic and paralinguistic value of the intonation variations. This expectation is related to the general assumption of the linguistic value of contour type, and the para-linguistic function of its concrete phonetic realization. The choice of contour would then be more related to, e.g., the type of sentence, while the pitch level and excursion size of the pitch movements would be more related to the speaker's emotional state.

In this light, Ladd et al. (1985 [24]) went on with the line of research described above, and this time, three experiments were conducted. Subjects judged the emotion conveyed in utterances in which F_0 range, type of pitch contour, and voice quality were systematically varied. Two separate rating forms were used, one for arousal-related states, the other for cognitively related attitudes. The type of pitch contour was processed in terms proposed by Ladd (1983 [23]). The first experiment aimed to assess the relative contribution of these three prosodic cues. The second was a replication of the first one as for generalizing results for F_0 range and type of contour. The third experiment tested whether pitch range variations have continuous or categorical effects on the perception of emotion. It was shown that F_0 range and voice quality had strong effects on the listeners' inference of the arousal-related state of the speaker, but also on the inference of cognitively related attitudes. The expectation that type of pitch contour would mostly affect the ratings of the cognitively related attitudes, could not be verified. In fact, the effect of the type of contour affected the ratings of emotion more then those of attitude. Findings concerning this point were also not conclusive in the second experiment. Moreover, and as could be expected, differences in F_0 range provoked continuous rather than categorical effects on judgments concerning the emotion of the speaker. Finally, the important conclusion could be drawn that the three prosodic cues F_0 range, voice quality, and type of pitch contour function independently of each other for conveying emotion and attitude in speech.

Another interesting way of considering this issue of linguistic and paralinguistic value of pitch variations in emotional speech is to conduct cross-cultural studies. Indeed, for the expression of emotion, prosodic phenomena show similarities in different languages, as well as differences from one language to another (van Bezooijen, 1984 [4]). However, relatively few cross-language studies focusing on emotional speech have been conducted. Tickle (1999 [41]) addresses methodological issues surrounding cross-cultural studies of speech conveying emotion. She distinguishes the two influences of biological factors, leading to the expectation of quasi-universal expression of emotions across cultures, and culturally determined factors, leading to expectations of differences in expression between cultures.

Finally, it seems appealing to enlarge the focus of studies interested in variability in speech in order to consider the incidence of socio-linguistic conventions on expressive speech. A broader range of meaningful speech variations would then be included in our studies of intonation. In this context, it also seems attractive to adopt the notion of 'expressive speech' instead of the notion of 'emotional speech', as this would better reflect our interest in processing expression and meaning.

3. Discussion

The use of a theoretical framework appeared to be very useful for processing intonational variations in a comprehensive manner. It enhances the degree of control of parameters under study, which is of benefit to the methodology. In this context, an approach involving the successive analysis of natural speech, the re-synthesis of speech allowing manipulations of natural speech, and the rule-based synthesis of speech, constitutes a valuable methodological background. However, the main advantage of considering data in the framework of a model of intonation seems to be that it makes it possible at the same time to make progress in modeling the variability conveying the expression of emotion in speech, and to check the validity of assumptions and models. It seems advisable to remind that intonation units differ, depending on the model. Units might be for instance tones, pitch movements, or phrase and accent commands. As different things are modeled, either targets, or whole pitch contours, comparisons across approaches are certainly not straightforward. May it be the case that the intonation units of the model used as framework for the investigation would appear not to be satisfying, or that the inventory of units would appear to be incomplete, this would affect the description of the data. Information in the F_0 curve that appears to be relevant but cannot be represented in the framework of the model, give indications that the model needs to be extended or adjusted.

Trying to model expressive speech forces us to model speech including all types of variations occurring in human speech. Emotional speech is an excellent example of speech containing a large variety of variations. These variations are of different types, i.e., variations in pitch, in speech tempo, rhythm, voice quality, loudness, articulation/pronunciation, and can be of a considerable size. Therefore, the study of emotional speech provides an exceptional opportunity of studying variability in speech. To what extent the variations are linguistic or paralinguistic in nature is an issue related to the expressive function of prosody. In the challenging context of the functionality of prosody, it is unclear to what extent the expression of emotion can be distinguished from the expression of meanings of linguistic nature such as the realization of a question or an exclamation. However, it seems desirable to distinguish between the expression of physiologically conditioned states and the expression of more cognitively related attitudes while extending the field of investigation to all meaningful speech variability.

The need to integrate the information about the emotional state of the speaker into the stream of information about the prosodic boundaries, the accents, their salience, and so forth, corresponds to ambitious aspirations in the field of speech technology. Although quantitative data will be necessary, merely carrying out analyses of speech followed by statistical analyses might not be the most rewarding approach in such an interdisciplinary field of research.

In the present paper, the type of contribution proposed by the tradition of intonational study to the study of the expression of emotion in speech was specified. The methodological considerations discussed should benefit studies in the enlarged context of the investigation of expressive speech. Summarizing, in the context of a model of intonation, and independently of whether use is made of a top-down or a bottom-up approach, it seems advisable to supplement production studies with perception studies. Indeed, complementary studies of production and perception are felt to be a necessary prerequisite for establishing the communicative significance of the investigated speech parameters. It also seems important to consider both the phonological information on the type of contour, and the phonetic information on the concrete pitch implementation of the F_0 curve. The investigation of type of pitch contour and detailed pitch implementation seems particularly promising when the experimental design is such as to allow the independent study of each effect, as well as their combined effect. This may lead us to make use of an orthogonal experimental design in the investigation of prosodic parameters.

Moreover, it is now rather well agreed upon that the speech variability corresponding to the expressiveness in the speech is not random and that a better understanding of this variability would be praiseworthy. Hence, it is assumed that including appropriate variability in synthetic speech would increase its naturalness, and its acceptance by human users. It then seems worthwhile to carry out experimental evaluations of naturalness in order to estimate the increase in naturalness obtained by taking prosodic variations conveying expressiveness into account, and to diagnose further lack of variability for specific uses. For now, the question remains whether the gain obtained by including supplementary relevant variability in speech would result in a more pleasurable interaction with TTS- systems, in an increased degree of acceptance of interaction with a machine, or in a reduced cognitive load on the system user, in comparison with an interaction with less variable synthetic speech.

Finally, non-vocal paralinguistic features such as co-speech gestures, posture, gaze, facial expression, and proximity changes are all relevant to expression. The term 'paralinguistic' can be used in the visual as well as the auditory modality. Considering this might constitute a motivation for a further extension of our interdisciplinary

field. That might lead us to study linguistic and paralinguistic aspects of prosody in a cross-modality field.

4. References

[1] Amir, N.; Ron, S.; Laor, N., 2000. Analysis of an emotional speech corpus in Hebrew based on objective criteria. *Proceedings of the ISCA Workshop on Speech and Emotion 2000: A conceptual framework for research*. Newcastle, Northern Ireland, pp. 29-33.

[2] Banse, R.; Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology* 70(3), 614--636.

[3] Beckman, M.E., 1997. Speech models and speech synthesis. In *Progress in speech synthesis*, J.P.H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirschberg (Eds.). Springer-Verslag, New-York, 185-209.

[4] Bezooijen, R.A.M.G. van, 1984. *The characteristics and recognizability of vocal expression of emotion.* Foris, Dordrecht, The Netherlands.

[5] Cahn, J.E., 1990. *Generating expression in synthesized speech.* Technical report, MIT Media Lab., Boston.

[6] Carlson, R., 1991. Synthesis: modelling variability and constraints. *Proceedings of Eurospeech 91.* Genova, Italy, Vol. 3, 1043-1048.

[7] Carlson, R.; Granström, B.; Nord, L., 1992. Experiments with emotive speech: acted utterances and synthesized replicas, *Proceedings ICSLP 92.* Banff, Alberta, Canada, Vol. 1, 671-674.

[8] Cosmides, L., 1983. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance* 9, 864-881.

[9] Di Cristo, A., 2000. Interpréter la prosodie. *Actes des XXIIIèmes Journées d'Etude sur la Parole.* Aussois, France, 13-29.

[10] Ekman, P., 1982. *Emotion in the human face*. Second edition. Cambridge University Press, New York.

[11] Fairbanks, G.; Pronovost, W., 1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs* 6, 87-104.

[12] Fienberg, S.E., 1980. *The analysis of cross-classified categorical data, second edition.* The MIT Press, Cambridge, Massachusetts.

[13] Frijda, N.H., 1986. *The emotions*. Cambridge University Press, Cambridge, England.

[14] Fujisaki, H., 1991. Modeling the generation process of F_0 contours as manifestation of linguistic and paralinguistic information. *Proceedings of the XIIth International Congress on Phonetic Sciences.* Aix-en-Provence, France, Supplement, 1-10.

[15] Grabe, E.; Gussenhoven, C.; Haan, J.; Marsi, E.; Post, B., 1997. Preaccentual pitch and speaker attitude in Dutch. *Language and speech* 41(1), 63-85.

[16] Gussenhoven, forthcoming. In *Prosodic Typology and Transcription: A Unified Approach*, C. Jun, S-A (Ed.). Oxford University Press.

[17] Hart, J. 't ; Collier, R. ; Cohen, A., 1990. *A perceptual study of intonation.* Cambridge University Press, Cambridge.

[18] Heuven, V.J. van, 1994. Introducing prosodic phonetics. In *Experimental studies of indonesian prosody*, C. Odé and V. J. van Heuven (Eds.). Semaian 9, Department of languages and cultures of South East Asia and Oceania, Leiden University, Leiden, 1-26.

[19] Higuchi, N. ; Hirai T. ; Sagisaka, Y., 1997. Effect of speaking style on parameters of fundamental frequency contour. In *Progress in speech synthesis*, J.P.H. van Santen, R.W. Sproat, J.P. Olive, J. Hirschberg (Eds.). Springer-Verslag, New-York, 417-428.

[20] Izard, C.E., 1977. *Human emotions*. Plenum Press, New York.

[21] Kitahara, Y.; Tohkura, Y., 1990. The role of temporal structure of speech in word perception and spoken language understanding. *Proceedings of the International Conference on Spoken Language Processing 90.* Kobe, Japan, Vol. 1, 389-392.

[22] Kitahara, Y.; Tohkura, Y., 1992. Prosodic control to express emotions for man-machine interaction. *IEICE Transactions on Fundamentals of Electronics, communications and computer sciences* 75, 155-163.

[23] Ladd, D.R., 1983. Phonological features of intonational peaks. *Language* 59, 721-759.

[24] Ladd, D.R.; Silverman, K.E.A.; Tolkmitt, F.; Bergman, G.; Scherer, K.R., 1985. Evidence for the independent function of intonation contour type, voice quality, and F_0 range in signalling speaker affect. *Journal of the Acoustical Society of America* 78, 435-444.

[25] Laver, J., 1995. The phonetic description of paralinguistic phenomena. *Proceedings of the XIIIth International Congress on Phonetic Sciences.* Stockholm, Sweden, Supplement, 1-4.

[26] Lieberman, P.; Michaels, S.B., 1962. Some aspects of fundamental frequency and envelope amplitude as related to emotional content of speech. *Journal of the Acoustical Society of America* 34, 922-927.

[27] Mozziconacci, S.J.L., 1995. Pitch variations and emotion in speech. *Proceedings of the XIIIth International Congress on Phonetic Sciences.* Stockholm, Sweden, Vol.1, 178-181.

[28] Mozziconacci, S.J.L., 1998. *Speech variability and emotion: Production and perception.* Ph.D. thesis, Technical University Eindhoven.

[29] Mozziconacci, S.J.L., 2001. Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction* 11, 297-326.

[30] Patterson, D.; Ladd, D.R., 1999. Pitch range modeling: linguistic dimensions of variation. *Proceedings of ICPhS 99.* San Francisco, USA, 1169-1172.

[31] Plutchik, R., 1980. *Emotion: a psychoevolutionary synthesis.* Harper & Row, New York.

[32] Protopapas, A.; Lieberman, P., 1997. Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America* 101(4), 2267-2277.

[33] Roach, P., 2000. Techniques for the phonetic description of emotional speech. *Proceedings of the ISCA Workshop on Speech and Emotion 2000: A conceptual framework for research.* Newcastle, Northern Ireland., 53-59.

[34] Rosis, F. de; Grasso, F., 2000. Affective natural language generation. In *Affective interactions,* A. Paiva (Ed.). Lecture notes in Artificial Intelligence, Vol. 1814, Springer-Verslag.

[35] Siegwart, H.; Scherer, K.R., 1995. Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi. *Journal of Voice* 9(3), 249-260.

[36] Scherer, K.R.; Ladd, D.R.; Silverman, K.E.A., 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America* 76 (5), 1346-1356.

[37] Scherer, K.R., 1991. Emotion expression in speech and music. In *Music, language, speech, and brain*, J. Sundberg, L. Nord, and R. Carlson (Eds.). Wenner-Gren Center International Symposium Series, MacMillan, London, 146-156.

[38] Scherer, K.R., 1996. Adding the affective dimension: a new look in speech analysis and synthesis. *Proceedings of the 4th International Conference on Spoken Language Processing.* Philadelphia, USA, Addendum, 20-23.

[39] Schlosberg, H., 1954. Three dimensions of emotion. *Psychological review* 61, 81-88.

[40] Schröder, M., 2000. Experimental study of affect bursts. *Proceedings of the ISCA Workshop on Speech and Emotion 2000: A conceptual framework for research.* Newcastle, Northern Ireland, 132-137.

[41] Tickle, A., 1999. Cross-language vocalisation of emotion: methodological issues. *Proceedings of ICPhS 99.* San Francisco, USA, 305-308.

[42] Williams, C. E.; Stevens, K.N., 1972. Emotions and speech: some acoustical factors. *Journal of the Acoustical Society of America* 52, 1238-1250.

[43] Zei, B., 2002. A place for affective prosody in a unified model of cognition and emotion. *This issue.*