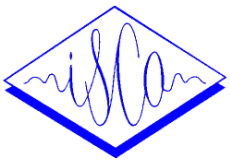

ITRW on
Speech and Emotion
Newcastle, Northern Ireland, UK
September 5-7, 2000

THE EXPRESSION OF EMOTION CONSIDERED IN THE FRAMEWORK OF AN INTONATION MODEL

Sylvie J.L. Mozziconacci [1,2]

[1]Phonetics Lab., Leiden University, PO Box 9515, 2300 RA Leiden, The Netherlands
[2]CREST, JST (Japan Science and Technology)

ABSTRACT

Studying intonational variation conveying the expression of emotion in speech is presented here as being particularly attractive when the pitch variation is represented in the theoretical framework of a model of intonation. Using such a framework while considering the meaning associated with these variations in terms of identification of emotion in speech facilitates a generalization of study results. Moreover, such an approach provides the opportunity to test the adequacy of the model for processing such extreme variations as the ones occurring in emotional speech. The model can be evaluated and could eventually be extended or fine-tuned. Ultimately, interpreting data in the framework of such a model provides a methodological background for investigating intonation phenomena conveying emotion in speech against the background of intonation phenomena relevant to speech communication. The approach proposed is discussed and illustrated by reporting how it has been applied, in particular to own work, using a specific model of intonation. Using this approach while investigating how prosody conveys meaning would stimulate us to consider the power of existing intonation models for modeling expressiveness. Hence, such an approach helps shedding light on functionality of intonation in terms of conveying meaningful information in the oral communication stream. From this point of view, studies on emotional speech contribute to developing and refining theories concerned with prosody.

1. INTRODUCTION

Communication is not merely an exchange of words. Linguistic elements, paralinguistic elements, non-verbal communicative elements such as facial expressions, co-speech gestures, non-speech sounds (clicks, sighs…) are part of the communication and all convey meaning. 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, his/her age and physical condition. Prosody can add information to the strict linguistic content of the message or even modify the meaning of the literal content. In this paper, the investigation of prosodic variation conveying the expression of emotion in speech will be discussed. While considering emotional speech, the focus will be on modeling intonation. Moreover, the intonation conveying emotion will be considered here in the context of speech synthesis. 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; Beckman, 1997).

Speech technology is a field in evolution. Retrospectively, the main concerns of this field were to synthesize intelligible speech, and to attain a high proportion of correct word identification in automatic speech recognition. As to speech synthesis, people only considered the use of speech technology in a rather limited set of situations, as the perspectives were more in line with development of reading machines than with the design of dialogue systems. With time and the development of our technical skills, our views on the use of these technologies have changed, and our expectations have grown towards diversified applications involving speech technology. In particular, the growth of Internet and the informal style of communication it allows has led to modifications of our wishes and motivations. Expectations towards speech technology now include dialogue systems allowing human-machine interactions. For adequate interactions, an interpretation of meanings is needed as well as an adaptation of the system response to the behavior of the human user. Robustness to different types of variations is expected from recognition systems. All together, there is an increased awareness that we will have to cope with speech variability. Correspondingly, the speech materials studied have changed over time, reflecting these increased expectations. At first, a rather "standard" type of speech has been studied. It was mostly read-aloud speech, often in the rather declarative style employed by a newscaster, while present studies more frequently investigate spontaneous speech and tend to involve a certain variety of speech styles, including, for instance, emotional speech. Moreover, synthetic speech is nowadays found to be readily intelligible. However, its lack of variability seems to remain the main obstacle to its more general use. Evaluations often point to the lack of proper prosody in synthetic speech, which seems to lead to uninvolved and unnaturally sounding speech. Especially now that synthetic speech is rather easily understandable, efforts are more often directed towards generating prosodic variation. Indeed, it is presumed that adding variation to synthetic speech in a proper way would enhance its naturalness and its acceptability.

Finally, if the study of speech is already considered an interdisciplinary field of research, considering the study of emotional speech, especially in the light of the present expectations towards speech technology only seems to enhance this interdisciplinary aspect. In fact, such a study constitutes a true meeting point of various disciplines. Moreover, the field of our interests has enlarged correspondingly to our ambitions towards speech technology applications. Numerous factors influence the production and perception of intonation. The

further restriction of the focus to the pitch variations in emotional speech does not prevent the topic from constituting a crossroad of diverse disciplines such as linguistics, psychology, neuro-psychology, phonetics, ethnology, cognition, artificial intelligence, acoustics, physiology, and information technology. Consequently, specialists of various disciplines are involved in this field of research, which is at the same time enriching and resulting in the coexistence of different approaches reflecting traditions of research in these different areas. Approaches are different, notions involved and parameters studied vary, material and measures differ, theories have different grounds, and even motivations for performing the studies are different. This tends to make comparisons among studies quite difficult. However, 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 intonational study to the study of the expression of emotion in speech.

2. INTONATION

Different prosodic cues are available to speaker and listener in order to encode and decode the full spoken message. This message does not only contain the verbal content but also additional information, among others information about the speaker's view, emotion, and attitude towards the topic, the dialogue partner, or the situation. All the information contributes to the interpretation of the message. Despite the fact that prosodic cues other than the intonational ones are involved in this process, we will, in this paper, only address intonation theory, and intonational cues conveying emotion in speech. First, main approaches to intonation, and types of intonation models in which we can attempt to represent relevant F_0 variations will be briefly presented. Systems for the transcription of intonation have also been developed and are available. Nevertheless, less attention has been paid to the meanings associated with pitch contours. The interest in processing speech variability and determining to what extent this variability is associated with meaning is still a relatively recent trend. A further discussion of this issue will be postponed till Section 4.

2.1. Intonation Models

Generally speaking, a distinction can be made between two main types of models of intonation both trying to account for those aspects of the F_0 curves that are relevant to speech communication. In the first type of models, F_0 curves are considered as the superposition of relatively fast pitch movements on a slowly declining line, i.e., the declination line or 'baseline'. Models of this type are called two-component models of intonation. One component represents the global aspect of the pitch curve; it represents the declination line and relates to pitch level. The other component represents the size of more local F_0 fluctuations and relates to pitch range. The second type of intonation models performs the modeling with a single component. A further distinction is that some models of intonation attempt modeling the whole pitch curves, while others model phonologically representative F_0 targets. Hence, models of each of these types are available. Gårding (1983) proposed to model whole contours by only using falling F_0 movements as components. Liberman and Pierrehumbert (1984) also proposed modeling intonation with a single component, but by means of F_0 targets, making use of the notion of downstep. Fujisaki (1991) proposed modeling whole contours with two components, while Pierrehumbert (1980) proposed a two-component model with F_0 targets. These are just examples of models available that illustrate the distinction proposed among model types.

Intonation models, independently of their type, are an expression of our need of theories of intonation as a basis for studying speech in a constructive way. They constitute a tool for representing and interpreting relevant data. Hence it should be noted that what is modeled is depending on the model itself. Therefore, depending on the interpretation in terms of one or the other particular model, differences could be observed between analyses of the same material. Results can lead us to confirm the usefulness of the approach or to refute assumptions underlying the model in question. In order to discuss actual instances of our considerations, a few illustrative studies are presented in Section 3. In Mozziconacci (1998), a specific model of intonation, i.e., the IPO model, is employed for the study of speech conveying emotion and attitude.

2.2. IPO Approach

The IPO approach is an experimental-phonetic approach to intonation ('t Hart et al., 1990). Perceptually oriented, it performs a data reduction, using perception as a filter in order to avoid modeling variations that are not relevant to perception, and therefore not relevant to communication. This reduction of data is realized by the stylization method. By this method, first, a simplification of the F_0 curve is made which contains all and only the perceptually relevant pitch movements in the utterance. Such a simplification, called a close-copy stylization of the original F_0 curve, is formed by the smallest possible number of straight-line segments in a linear time versus log frequency (or ERB) plot. When resynthesized, the close-copy is perceptually equal to the original. It forms a starting point for making standardized contours. Such contours need not sound exactly identical to the original; they may be audibly different, but can still be considered 'melodically equivalent' to the original. They can therefore be considered as successful imitations of the original.

The IPO model of intonation is a two-component model treating whole contours as combinations of straight-line segments, each segment corresponding to a single pitch movement. In this model, the end point of the declination line represents the pitch level, while the excursion size of the pitch movements represents the pitch range. In the most basic version of the model, the excursion size of the pitch movements is considered to be constant throughout the utterance, so that pitch contours could also be described with a lower declination line, or baseline, and an upper declination line, or topline, between which the pitch movements are realized. The overall excursion size of the pitch movements then equals the distance between the lower and the upper declination line.

Additionally, for a few languages, a "grammar" of intonation has been developed. For these languages, an inventory has been made of standard acoustic specifications for each perceptually

distinct *pitch movement*. Pitch movements are characterized by their timing in the syllable, their spread over one or several syllables, and their size relative to the topline, e.g., full or half. A functional characteristic is whether the pitch movement may or may not lend prominence to a syllable. In the case of Dutch (’t Hart et al., 1990), five rises and five falls were specified as in Table1. Additionally, ‘0’ and ‘∅‘ stand for the pitch level on the lower and the upper declination lines, respectively, and ‘&’ links two pitch movements occurring on a single syllable. According to the theory, these pitch movements combine into *configurations*, which in their turn combine into *pitch contours*. The combination rules of these pitch movements into configurations and into pitch contours are independent of the specific excursion size of the pitch movements, and constitute a grammar of intonation. This grammar defines an inventory of legal sequences of pitch movements, which in principle is unlimited. At the highest level of description, it is presumed that these different pitch contours in unlimited number are manifestations of a finite number of basic intonation patterns. One of the remaining questions is how the speaker makes a choice for one of these legal sequences of pitch movements. The grammar makes it possible to generate pitch contours from specifications of accent places, and pitch movement or configuration labels, and to analyze surface-phonetic F_0 curves into pitch contours.

Label	Movement	Timing	Prominence lending	Other specification
1	rise	early	yes	
2	rise	very late	no	
3	rise	late	yes	
4	rise		no	extent: various syllables
5	rise	early	yes	half rise, i.e., overshoot
A	fall	late	yes	
B	fall	early	no	
C	fall	very late	no	
D	fall		no	extent: various syllables
E	fall	early	yes	half size

Table 1: Specifications of pitch movements in Dutch.

3. INTONATION CONVEYING EMOTION IN SPEECH

Various studies stated the importance of intonation as a medium for expressing emotion in speech (e.g. Williams and Stevens, 1972; Cosmides, 1983; Cahn, 1990; Kitahara and Tohkura, 1990). In the present section, a few specific examples of approaches to intonation conveying emotion are described, enhancing the contribution of and/or to general knowledge and assumptions concerning intonation.

Investigation in the framework of a model of intonation Mozziconacci (1998) investigated intonation in production and perception of Dutch speech conveying six emotions or attitudes: joy, boredom, anger, sadness, fear, and indignation, against neutrality as a reference. The issue of the identifiability of the “emotions” was addressed first. The study involved perception tests with natural speech, manipulations of natural speech through analysis-resynthesis, and synthetic speech, successively. In the first stage of the study, speech material successfully conveying emotion in speech was selected on the basis of a perception test. Optimal values were sought at utterance level, for the global parameters: pitch level, pitch range, and speech rate. These values were derived for the generation of emotional speech from a neutral utterance, and perceptually tested in re-synthesized speech and in synthetic speech. Then, in a production study, the speech of three speakers was analyzed, involving, at the utterance level, the global measures pitch range, pitch level, and speech rate. A more local analysis, taking variations within utterances into account, involved the prosodic features: relative height of the pitch accents, and final lowering. The F_0 curves of the emotional speech produced by the three speakers were analyzed by means of pitch measurements at anchor points chosen in the utterances. Final lowering and relative height of pitch-accent peaks appeared to be two major sources of deviation between the rule-based pitch curves – tuned as for pitch level and range – that were synthesized according to the IPO model of intonation, and the F_0 curves actually realized by the speakers. The relevance of these deviations was perceptually investigated. Another part of the local analysis was concerned with the shape of the pitch curves. The stylized F_0 curves were labeled in terms of pitch contours, using the intonation grammar for Dutch (’t Hart et al., 1990). Configurations of pitch movements realized by the speakers in the initial and the final parts of utterances were then considered separately, and the distribution of the configurations of pitch movements over the various emotions was investigated. The perceptual relevance of the choice of type of pitch contour for the identification of emotion was tested in an experiment, and a cluster analysis was run on the results.

Parameter values best used for conveying the emotions studied are summarized in Mozziconacci (1998), and are beyond the scope of the present paper. However, in addition to the results directly concerned with the expression of emotion in speech, the study also sheds light on a few more general issues, which will be discussed here. Indeed, a study considering speech variations as extreme as the ones occurring in the expression of emotion, is a source 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 the consideration of how to modify the model can contribute to our understanding of speech variation. In this study, analysis, synthesis, and manipulation by means of analysis-resynthesis were carried out within the framework of the IPO model of intonation, which allowed for controlling parameters, enhancing the systematic aspect of procedures, and testing the adequacy of the model for processing emotional speech.

Percentages correct identification of emotion are reported in Table 2 for three types of stimuli, used in three different experiments: close-copy stylization of natural speech, speech generated by manipulating pitch level and pitch range of natural

	neu	joy	bor	ang	sad	fea	ind	mean
Close-copy	85	62	92	32	97	60	85	73
Manip. of pitch level and range	67	72	85	42	75	42	77	66
Rule-based diphone synthesis	83	62	94	51	47	41	68	63

Table 2: Percentage correct identification of the emotions.

utterances, and 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, 1995), i.e., here 14.3, the results can be considered acceptable, and the IPO model considered adequate for our purpose.

When studying speech variability, estimating pitch level and pitch range is frequently done by means of mean and standard deviation of F_0. Such crude measures must be expected to obscure a substantial part of the variation present in the speech material, and do not provide any information concerning the linguistically relevant variation. Their frequent use is probably due to the fact that they are easy to obtain, and that their common use facilitates comparison of results across studies. Moreover, the notions of pitch level and range they stand for correspond to parameters in most synthesizers. Mozziconacci (1998) discussed two ways of estimating pitch level and pitch range. One estimation, strictly data oriented, was based on the mean and standard deviation of F_0, respectively. The other estimation, model-based, was involving the end point of the baseline and the difference between baseline and topline, respectively. It appeared that, though perhaps not very accurate, the crude measures remain quite informative.

For an investigation of pitch variations within utterances, the production data were first represented as F_0 targets at anchor points, using a tonal approach, and then described within the IPO's model of intonation. It appeared that some details that were observed with the tonal approach, involving pitch measurements at anchor points within the utterances, could not be captured in this model. These details concern the relative height of the accent peaks and the final lowering, and provide valuable supplementary information to the mean F_0 and its standard deviation when estimating pitch level and pitch range by means of these global measures. However, a perceptual evaluation of the relevance of these details showed that they are not very important for the expression of most emotions in speech. Therefore, the difficulty to represent this detailed information within the model should not be considered a major problem. It just means that the model provides a simplification of the pitch phenomena on the basis of perceptual relevance, and does not undermine its adequacy for describing speech, even when the speech involves a wide range of variations. Moreover, it was speculated that other two-component intonation models would also be adequate. Indeed, Higuchi et al. (1997) also carried out an experiment seeking optimal values for pitch level, pitch range, and speech rate. This time, the framework of Fujisaki's model of intonation (1991) 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.

For a description of the course of F_0 in time, an intonation labeling was carried out in terms of the IPO intonation grammar for Dutch by 't Hart et al. (1990). Sequences of labels considered legal in this grammar could be attributed to the large majority of concrete combinations of pitch movements. The contours were not equally distributed over the emotions. In order to test the effect of contour type on identification of emotion, a perception experiment was run, testing eleven contours in combination with each of the seven optimal sets of values for pitch level, pitch range, and speech rate, i.e., one per emotion. As a result, relationships could be established between contours and emotions. For some contours, the identification of some emotions could increase, while for others, it could decrease. However, the '1&A' configuration, i.e., a prominence-lending rise-fall, which was the most frequently used contour and was regularly produced in all seven emotions showed to be a contour that can potentially convey all emotions, when it is desirable, for instance for methodological reasons, that no variability is introduced by the realization of different contour types. Hence, no specific pitch contour appeared to be strictly necessary for signaling emotion in speech. Furthermore, a cluster analysis of the contours that was performed on the perception data, showed that predominantly the final configuration of the contour plays a role in conveying emotion. The clustering reflected perceptual distinctions among contour types. For instance, the contours '1&A 1&A', '1B 1&A', '1D 1&A', and '12 1&A', form one cluster, while the contours '1&A 3C', '1B 3C', and '1D 3C' form a second cluster.

It appeared that the F_0 curves of natural emotional speech could be described using this approach, yielding the adequacy of the approach for describing F_0 curves conveying emotion. An extension of the intonation grammar for Dutch was not necessary. The distinctive features presented in this grammar appear to be sufficient. Pitch contours generated within this approach contributed to conveying emotion. Some percentage of correct identification of emotion obtained in the perception experiment are reported in Table 3. For generating the stimuli used in this experiment, contours were transplanted onto a single carrier utterance per text. This carrier had the voice quality and temporal features of the neutral utterance. Percentages are low, but it should be considered that stimuli were not prepared as instances of any intended emotion, but rather as instances of all possible combinations of contour type and 'pitch', i.e., pitch level and pitch range. Results in the successive rows show the

	neu	joy	bor	ang	sad	fea	ind	mean
Condition 1	37	18	41	9	8	19	24	22
Condition 2	46	10	48	10	0	17	19	21
Condition 3	56	27	19	27	33	10	4	25
Condition 4	56	35	52	23	19	25	52	37

Table3: Percentage correct identification of the emotions per condition
C1. all contours in combination with optimal pitch (per emotion).
C2. only the '1&A 1&A' contour, in combination with optimal pitch.
C3. optimal contour (per emotion), in combination with neutral pitch.
C4. optimal contour, in combination with optimal pitch.

percentage identification. Conditions 1 and 2 test the effect of optimal 'pitch' per emotion, averaged over all contour types, and averaged only over '1&A 1&A', respectively. Condition 3 tests the effect of contour type, and Condition 4 tests the combined effect of contour type and 'pitch'. In comparison with each other, results reflect the significant main effects of pitch and contour type on subjects' responses. Moreover, assuming a chance level of 14% corresponding to 7 emotion categories, the contribution of optimal implementation of pitch is limited to +8%, i.e., raising the score to 22%. The contribution of phonological choice of contour type is greater, i.e., 11%, raising the score to 25% correct identification. The effects of phonological choice of 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 37% correct identification obtained in condition 4. The slight superadditivity is not significant. Moreover, the results reflect the independent effects of the implementation of pitch level and range, and the selection of contour type on the identification of emotion. Results show that it is the phonological choice of contour that primarily determines the emotion perceived, while the phonetic implementation in terms of pitch level and range is of secondary importance.

Moreover, if the IPO approach to intonation appears adequate for conveying emotion in speech, one can wonder whether making use of another approach to intonation would have led to similar results. In a different approach to intonation, such as an auto-segmental approach, a different type of model of intonation would be employed. However, the choice of the IPO two-component model does not imply that another model of intonation would have been less suited for use in this study. A simple comparison between intonation labeling in the IPO approach and in the auto-segmental phonological approach is, however, not obvious. Both approaches provide a labeling, but at different levels. There is, for example, no universally agreed way of transcribing the three configurations of pitch movements '1&A', '1B', and '1D' into their auto-segmental counterparts. These configurations, which all correspond with the first accent produced in the utterances, could correspond, if simplification is allowed, with a bitonal H*L accent in the auto-segmental description of intonation. Since all pitch contours that were different only in this first part were grouped into the same clusters, it is to be expected that, using an auto-segmental transcription system, even if it makes different distinctions in the group of '1&A', '1B', and '1D' configurations, will lead to the same general conclusions.

Finally, the 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 for this study. Complementary studies of production and perception were felt to be a necessary prerequisite for establishing the communicative significance of the investigated speech parameters. This need for a combination of both production and perception studies, in order to make progress in understanding emotional vocal communication, was already expressed, for instance, by Scherer (1991). A certain type of variation can be perceptually relevant, while not systematically showing up in the production data. Indeed, the efficient use of a particular parameter can allow the relaxation of another one. That other parameter can be potentially relevant. On the other hand, in a perception study, the perceptual relevance of a cue observed in a production study may be obscured by the experimental set-up or the effect of another cue. Alternatively, the cue can simply be relevant to an aspect of the communication process other than the one studied. However, the correspondence of results obtained in the production and perception studies discussed here firmly establishes the communicative importance of the parameters being studied.

Testing two approaches Another very interesting way of inheriting the benefit of intonation theories, is to use emotional speech for testing assumptions that are explicitly or implicitly underlying such theories.

Scherer et al. (1984) 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 and information on the strictly linguistic content of utterances function independently of each other. 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 which emotion was conveyed (using a rating scale). 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) 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. It was also 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 sociocultural and linguistic conventions.

Ladd et al. (1985) went on with this line of research. 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). 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.

4. INTONATION CONVEYING EMOTION VERSUS EXPRESSION AND MEANING

Considering intonation data concerning emotional speech within the theoretical framework of approaches to intonation made quite clear that positioning the variations conveying emotion in speech among all variations contributing to oral communication is largely an unsolved issue. The variety of labels used in studies focussing on intonation in 'emotional speech' reflects this difficulty. Typical labels include joy, sadness, anger, fear, disgust, boredom, surprise, interest, reproach, doubt, disappointment, politeness, friendliness, rudeness, suspiscious irony, incredulous question, interrogation, and exclamation. Studies in this field mainly focus on the prosodic phenomena accompanying the verbal expression in speech and not on the psychological side of the matter. In this context, the term 'emotion' covers a large variety of expressions including notions such as emotion, attitude, intention, feeling, and even sentence type. Despite the different lists of emotions that have been proposed by different authors (e.g., Izard, 1977; Plutchik, 1980; Ekman, 1982; Frijda, 1986), and sometimes successively by the same author, there is no commonly accepted definition and taxonomy of emotion. This does not necessarily constitute an insurmountable methodological difficulty, as empirically based lists of relevant notions can be used, either involving notions considered useful in the context of verbal user-system interactions, or based on the ability of subjects to rely on empirically based notions. However, the lack of definition and taxonomy of emotion remains an issue, as 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 wish of enlarging the scope of our studies seems to relate to an underlying desire of comprehending the functionality of prosodic cues. Indeed, prosodic parameters fulfil various functions such as structuring the discourse, providing cues on turn-taking in dialogue, lending prominence, indicating the phrasing of sentences, contributing to speaker identity, and conveying the expression of emotion and attitude. Some of these functions are traditionally considered linguistic, and others paralinguistic in nature. However, for the expression of emotion, prosodic phenomena show similarities in different languages, as well as differences from one language to another (van Bezooijen, 1984). Although the prosodic function of conveying the expression of emotion seems to involve both a linguistic and a paralinguistic component (e.g. Laver, 1995), it is frequently considered a paralinguistic function, despite the doubts emitted on the subject (van Heuven, 1994). However, even if, as in the view developed by Ladd et al. (1985), the expression of emotion is considered merely a paralinguistic issue, the interaction with the linguistic content still requires further investigation in the context of the interpretation of the general meaning of the message. An extension of the field of variations under study would also involve considering the incidence of socio-linguistic conventions on expressive speech. Moreover, in the context of speech technology, the information conveying emotion has to be integrated with the remainder of the linguistic information, for instance when computing prosody for Text-To-Speech systems. Hence, it seems appealing to include a broader range of meaningful speech variations in our studies of intonation, and to adopt the notion of expressive speech reflecting our interest in processing expression and meaning.

In the context of such an extension of the field of investigation involving the whole meaningful speech variability, we are confronted with the great difficulty in distinguishing acoustic correlates of arousal from those of cognitive attitude exposed in Ladd et al. (1985). In this study, distinguishing these two types of expression appeared useful in reaching conclusions. Moreover, although Mozziconacci (1998) did not explicitly draw a distinction between attitude and emotion, both notions were included in the investigation. Similar conclusions have been reached in Ladd et al. (1985), and Mozziconacci (1998). Indeed, in both studies, it appeared that prosodic features function independently of each other. No interactions were found either between pitch, i.e., pitch level and range, and type of contour (Mozziconacci, 1998), or between voice quality and pitch (Ladd et al., 1985). This is an important conclusion as it justifies the independent study of these cues in expressive

speech. Converging evidence was also found that both approaches tested by Ladd et al. (1985) are needed in this type of investigations. Indeed, emotions are conveyed by pitch, i.e. pitch level and range, and also conveyed by type of contour. Additionally, it has been shown that using both features, i.e., pitch and contour type yields better results.

However, the formulation used by Ladd et al. (1985) in terms of scalar versus categorical features raises a few questions. In this study, a continuous effect of scalar parameters on perception was found in the ratings concerning the force of the emotion. However, Mozziconacci (1998) found that the efficient use of a specific parameter could allow the relaxation of another one, which seems to induce that the force of the emotion does not need to be associated to a higher value for a scalar parameter, i.e., parameter corresponding to phonetic implementation. Another point relates to the assumption that scalar variables would be appropriate for conveying physiologically based emotional states, while cognitively related attitudes would be best conveyed making use of categorical variables, i.e., variables corresponding to a phonological choice such as the choice of pitch contour. Results on this issue were not conclusive in Ladd et al. (1985). In Mozziconacci (1998), the effect of the type of contour appeared indeed to be the strongest for the identification of indignation and neutrality – categories that are not strictly considered as emotions, but are rather cognitively related –. However, both studies have shown that the type of contour is relevant to conveying emotion as well as attitude in speech. These issues obviously require further investigation.

Furthermore, the interpretation of intonation in terms of meaningful information conveyed by prosodic cues has already given rise to quite some interest. Uldall (1972) has investigated the interpretation of various contour types, using different sentence types and ratings in terms of emotion. Caspers (1997) tested the meaning associated with different types of pitch contour in Dutch, and found that different types of contour expressed different intentions. Haan et al. (1997), in a study of intonation concerned with declarativity and interrogativity, have shown that not only the shape of pitch contours is relevant in order to mark type of question or declaration; scalar features, i.e., pitch level and pitch range, also constitute acoustic correlates of the type of sentence investigated.

Finally, 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.

5. 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. However, the main advantage 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 models. It would be cautious to keep in mind 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. It should be noted that if the intonation units of the model used as framework for the investigation would not be satisfying, or that the inventory of units would be incomplete, it 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 us 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. This issue is related to the current assumption that there is a correspondence between intonational categories and interpretative categories.

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, as it is assumed that including variability in synthetic speech would increase its naturalness and its acceptance by human users, it seems worthwhile to carry out experimental evaluations of naturalness in order to estimate the increase in naturalness obtained by taking prosodic variations into account, and to diagnose further lack of variability for specific uses. For now, the question remains whether the gain obtained by including more variability in speech would result in a more pleasurable interaction with TTS- systems, in an increase of the degree of acceptance of interaction with a machine, or in a reduced fatigability in comparison with an interaction with less variable synthetic speech.

Finally, reminding that non-vocal paralinguistic features such as co-speech gestures, posture, gaze, facial expression, and proximity changes are all relevant to expression, and that the term ‘paralinguistic’ can be used in the visual as well as the auditory modality might constitute a motivation for a further extension of our interdisciplinary field. That might lead us to study the linguistic and paralinguistic aspects of prosody in a cross-modality field.

6. ACKNOWLEDGMENTS

The experimental studies given as examples of my own work have been conducted at IPO. Additionally, I want to sincerely

thank Vincent van Heuven and Dik Hermes for their useful comments on a draft version of this paper.

7. REFERENCES

1. Beckman, M. E. (1997). “Speech models and speech synthesis”. In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.) *Progress in speech synthesis*, Springer-Verslag, New-York, 185-209.
2. Bezooijen, R. A. M. G. van (1984). *The characteristics and recognizability of vocal expression of emotion.* Foris, Dordrecht, The Netherlands.
3. Cahn, J. E. (1990). *Generating expression in synthesized speech.* Technical report, MIT Media Lab., Boston.
4. Carlson, R. (1991). Synthesis: modelling variability and constraints, *Proceedings Eurospeech’91, Genova, Italy, 3*, 1043-1048.
5. Caspers, J. (1997). Testing the meaning of four Dutch pitch accent types, *Proceedings Eurospeech’97, Rhodes, Greece, 2*, 863-866.
6. Cosmides, L. (1983). “Invariances in the acoustic expression of emotion during speech,” *Journal of Experimental Psychology: Human Perception and Performance, 9*, 864-881.
7. Fujisaki, H. (1991). Modeling the generation process of F_0 contours as manifestation of linguistic and paralinguistic information. *Proceedings XIIth ICPhS, Aix-en-Provence, France, supplement*, 1-10.
8. Gårding, E. (1983). “A generative model of intonation”. In: A. Cutler and D. R. Ladd (Eds.) *Prosody: Models and measurements*. Springer-Verlag, Berlin, 11-25.
9. Haan, J., van Heuven, V., Pacilly, J., and van Bezooijen, R. (1997). Intonational characteristics of declarativity in Dutch: a comparison. *Proceedings ESCA workshop on intonation: theory, models and applications, Athens, Greece*, 173-176.
10. Hart, J. ’t, Collier, R., and Cohen, A. (1990). A perceptual study of intonation. Cambridge University Press, Cambridge.
11. Heuven, V. J. van (1994). “Introducing prosodic phonetics”. In: C. Odé and V. J. van Heuven (Eds.) *Experimental studies of indonesian prosody*, Semaian 9, Department of languages and cultures of South East Asia and Oceania, Leiden University, Leiden, 1-26.
12. Higuchi, N., Hirai T., and Sagisaka, Y. (1997). “Effect of speaking style on parameters of fundamental frequency contour”. In: J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg (Eds.) *Progress in speech synthesis*, Springer-Verslag, New-York, 417-428.
13. Kitahara, Y., and Tohkura, Y. (1990). The role of temporal structure of speech in word perception and spoken language understanding, *Proceedings ICSLP 90, Kobe, Japan, 1*, 389-392.
14. Ladd, D. R. (1983). “Phonological features of intonational peaks”. *Language, 59*, 721-759.
15. Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergman, G., and 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.
16. Laver, J. (1995). The phonetic description of paralinguistic phenomena. *Proceedings XIIIth ICPhS, Stockholm, Sweden*, 1-4.
17. Liberman, M. Y., and Pierrehumbert, J. (1984). Intonational invariances under changes in pitch range and length. In: M. Aronoff anf R. T. Oehrle (Eds.) *Language sound structure: Studies in phonology presented to Morris Halle*. MIT- Press, Cambridge 157-233.
18. Mozziconacci, S. J. L. (1998). *Speech variability and emotion: Production and perception.* Ph.D. thesis, Technical University Eindhoven.
19. Pierrehumbert, J. (1980). *The phonetics and phonology of English intonation.* Ph.D. thesis, MIT.
20. Siegwart, H., and 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.
21. Scherer, K. R., Ladd, D. R., and Silverman, K. E. A. (1984). “Vocal cues to speaker affect: Testing two models,” *Journal of the Acoustical Society of America, 76 (5)*, 1346-1356.
22. Scherer, K. R. (1991). “Emotion expression in speech and music”. In: J. Sundberg, L. Nord, and R. Carlson (Eds.), *Music, language, speech, and brain*, Wenner-Gren Center International Symposium Series, MacMillan, London, 146-156.
23. Uldall, E. (1972). “Dimensions of meaning in intonation”. In: Bolinger, D. (Ed.), *Intonation*, Penguin, Harmondsworth, England, 250-259.
24. Williams, C. E., and Stevens, K. N. (1972). “Emotions and speech: some acoustical factors,” *Journal of the Acoustical Society of America, 52*, 1238-1250.