The role of intonation in emotional expressions

Banziger Flykt, Tanja; Scherer, Klaus R.

How to cite

BANZIGER FLYKT, Tanja, SCHERER, Klaus R. The role of intonation in emotional expressions. In: Speech Communication, 2005, vol. 46, n° 3-4, p. 252–267. doi: 10.1016/j.specom.2005.02.016

This publication URL: https://archive-ouverte.unige.ch/unige:96410
Publication DOI: 10.1016/j.specom.2005.02.016

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 46 (2005) 252–267
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The role of intonation in emotional expressions

Tanja Bänziger *, Klaus R. Scherer

Department of Psychology, FAPSE, University of Geneva, 40 bv du Pont-d'Arve, 1205 Genève, Switzerland

Received 28 August 2004; received in revised form 10 January 2005; accepted 4 February 2005

Abstract

The influence of emotions on intonation patterns (more specifically F0/pitch contours) is addressed in this article. A number of authors have claimed that specific intonation patterns reflect specific emotions, whereas others have found little evidence supporting this claim and argued that F0/pitch and other vocal aspects are continuously, rather than categorically, affected by emotions and/or emotional arousal. In this contribution, a new coding system for the assessment of F0 contours in emotion portrayals is presented. Results obtained for actor portrayed emotional expressions show that mean level and range of F0 in the contours vary strongly as a function of the degree of activation of the portrayed emotions. In contrast, there was comparatively little evidence for qualitatively different contour shapes for different emotions.

Keywords: Emotion; Expression; Intonation; Pitch-contour; Expressive-speech

1. Introduction

This paper examines the contribution of intonation to the vocal expression of emotions. Over the past decades, this question has been addressed by many authors from different research backgrounds and is still a matter of sustained debate. A tradition, emerging from the linguistic approach to the study of intonation contours, has claimed the existence of emotion specific intonation patterns (e.g. Fonagy and Magdics, 1963). However, the evidence offered for this notion consists mostly of selected examples rather than of empirical examination of emotional speech recordings. Efforts to describe/analyze the intonation of actual emotional expressions have been limited by the use of simplified descriptors, such as measures of overall pitch level, pitch range or overall rise/fall of pitch contours. Some authors have directly questioned the existence of emotion specific intonation patterns. Pakosz (1983), for instance, claimed that

* Corresponding author. Tel.: +41 22 379 92 07; fax: +41 22 379 92 19.
E-mail addresses: tanja.banziger@pse.unige.ch (T. Bänziger), klaus.scherer@pse.unige.ch (K.R. Scherer).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.016

intonation only carries information about the level of emotional arousal. In this perspective, elements of the context in which the expressions are produced and/or information carried by other channels (typically facial expressions) are required to disambiguate specific emotion categories.

In the following paragraphs, more details on (1) the linguistic approach to the description/analysis of intonation and (2) some results obtained on the basis of empirical analysis of emotional speech are reviewed. The limits of those approaches for the study of the intonation of emotional speech are introduced; and, finally, the approach used in the study presented in this paper will be outlined.

1.1. The linguistic approach to the description/analysis of intonation

Various definitions of concepts such as *intonation* or *prosody* have been proposed by authors working on the analysis and description of nonverbal features of running speech. Cruttenden (1986, pp. 2–3), proposed the following definition:

"The prosody of connected speech may be analysed and described in terms of the variation of a large number of prosodic features. There are, however, three features which are most consistently used for linguistic purposes either singly or jointly. These three features are pitch, length, and loudness. [...] Pitch is the prosodic feature most centrally involved in intonation and it is with this feature that I shall be principally concerned in this book."

As in the above citation, the definition of the term *intonation* generally includes aspects related to *pitch*, *length* and *loudness*; whereas, somewhat paradoxically, most authors focusing on *intonation* described and analyzed essentially *perceived pitch contours*. Transcriptions of pitch contours were first developed to account for linguistic functions of intonation, often with a didactic purpose. A great variety of transcription systems have been proposed over time. Thirty years ago, Léon and Martin (1970, pp. 26–32) distinguished, for instance, six forms of pitch transcription, including "musical transcriptions", "transcriptions of patterns of intonation" and "transcriptions representing levels of intonation". More recently, different models have been proposed for the linguistic analysis and description of intonation (perceived pitch). A broad distinction can be made between *tone sequence models* (such as the Tones and Break Indices system, ToBI, Silverman et al., 1992)—which describe pitch as a sequence of (high/low) tones on specified targets—and *superpositional models*—which define the overall pitch contour as the superposition of hierarchically ordered components.

The most prominent superpositional model (Fujisaki, 1988), includes two components: a phrase component (i.e. a contour defined at the phrase level) and an accent component (i.e. local excursions superposed to the phrase contour). Superpositional models allow to account for phenomena such as global pitch declination or anticipation effects in the production of overall pitch contours. But, independently of the number and the quality of the components included, superpositional models remain relatively abstract. With respect to actual pitch contours, one or more component(s) need to be fixed according to a set of more or less arbitrary rules, allowing to define the other component(s). Furthermore, relationships between those relatively abstract components of pitch contours and linguistic or paralinguistic functions are difficult to specify.

In recent years, tone sequence models have been more extensively used than superpositional models for the description and analysis of linguistic intonation (perceived pitch). ToBI (Tones and Break Indices system, Silverman et al., 1992)—a pitch transcription system originally derived from the tone sequence model developed for the intonation of English by Pierrehumbert (1980)—has been adapted and used extensively for the description of perceived pitch in several languages. In this coding system contrastive tone values (high/low) are attributed to linguistically defined targets. Relative pitch levels (tones) are allocated to accented syllables (pitch accents) and to intonative boundaries (phrasal tones, final boundary tones).

Linguistic models of intonation rely, more or less explicitly, on linguistic segmentations of the speech flow and their primary purpose is to

describe linguistic functions of intonation. To our knowledge, with one exception,[1] such models have not been systematically applied to the description and analysis of actual corpuses of emotional expressions. Accordingly, the possibility for those models to account for variations of intonation related to emotional expressions remains largely untested.

Nonetheless, the strongest claims supporting the existence of emotion specific intonation patterns originate in the linguistic approach to the description of intonation contours. Various authors (e.g. Fonagy and Magdics, 1963; Halliday, 1970; O'Connor and Arnold, 1973)—using different transcription models—proposed descriptions of pitch contours for specific emotions. Fonagy and Magdics' approach (1963) provides a good illustration of this approach. They described perceived pitch contours using sequences of tones on a musical score for various utterances corresponding to different emotional situations. The French utterance "*Comme je suis heureuse de te voir! Je ne pensais pas te rencontrer!*" is for instance reported to illustrate a typical joyful contour. Frick (1985) voiced several objections relatively to this kind of approach. In particular, he pointed out that the verbal content of the utterances used as examples often carries the meaning that the pitch contour is supposed to convey (as in the example borrowed from Fonagy and Magdics) and that the emotional impression a reader derives from the example is likely to arise from the implicit addition of prosodic elements unspecified in the pitch contour transcriptions provided by the author(s).

Empirical studies on the contribution of intonation to the communication of emotional meaning have been largely carried out independently of the models (or transcriptions) proposed in the research tradition described above. An overview of the evidence gathered on this issue is presented in the following section.

[1] Mozziconacci (1998) described emotional (acted) speech using a set of contours specified in the intonation model developed for Dutch by t'Hart et al. (1990).

1.2. Empirical studies on the contribution of intonation to the communication of emotion

A large number of studies have investigated vocal correlates of emotional expressions (for recent reviews see Juslin and Laukka, 2003; and Scherer, 2003). A common finding in those studies is that portrayed emotions influence global descriptors of F0, such as average F0, F0 level or F0 range. Reviews in this field show that portrayed emotions also have an effect on other broad descriptors of intonation, in particular measures derived from the acoustic intensity contour and measures related to the relative duration of various speech segments. Comparably few studies have attempted to describe F0 contours for different emotional expressions. In their review, Juslin and Laukka (2003) examined 104 studies concerned with the vocal communication of emotion, 77 studies reported acoustic descriptions for various expressed and/or perceived emotions, 69 studies used overall descriptors or manipulations of F0, and only 25 studies included a description of F0 contours or an attempt to influence emotional attributions through the systematic manipulation of F0 contours. When descriptions of F0 contours are provided they often come down to global "rising" or "falling" of the overall contour shape. According to Juslin and Laukka's review, rising contours were reported in 6 out of 8 studies for anger expressions, in 6 out of 6 studies for fear expressions, and in 7 out of 7 studies for joy expressions. Falling contours were reported in 11 out of 11 studies for sadness expressions and in 3 out of 4 studies for tenderness expressions.

This synthesis of the results described in the literature shows that studies reporting empirical results regarding contour shapes for emotional expressions are relatively scarce. It also reflects both the over-simplification and the heterogeneity of the descriptions used to characterize emotional intonation contours. The descriptions provided are often very rudimentary (such as global rise/fall) and more specific aspects reported in different studies can hardly be compared between studies. However, evidence for the participation of intonation (pitch/F0 contours) in the communication of emotional meaning can be derived from studies

that have not attempted to describe specific contours for specific emotions, but primarily tried to assess the respective importance of intonation and voice quality (voice timbre) for the communication of emotional meaning. Starting in the early sixties, several authors tried to separate the contribution of voice quality and of intonation by using various signal manipulation techniques (see for example Scherer et al., 1985). The following paragraphs describe the results obtained by two studies (Ladd et al., 1985; Lieberman and Michaels, 1962) demonstrating that isolated aspects of intonation can contribute to the communication of emotional meaning and two further studies (Scherer et al., 1984; Uldall, 1964) indicating that intonation patterns might influence emotional or attitudinal attributions in combination with the linguistic content of the expressions.

Lieberman and Michaels (1962) used expressions corresponding to eight emotional modes,[2] 85% of the recorded expressions were correctly identified by a group of listeners. The authors extracted the F0 contours of the original expressions and resynthesized them on a fixed-vowel. The proportion of "emotional modes" correctly identified dropped in this condition, but 44% of the expressions synthesized with the copied F0 contours were still correctly identified. When smoothing the F0 contours with a 40 ms time constant, this proportion further dropped to 38%; 100 ms smoothing reduced the recognition rate to 25%. This study indicates that F0 fluctuations can carry emotional meaning, independently from amplitude variations and voice quality. It also shows that short-time variations of F0 contours might be of importance in this process. Lieberman and Michaels (1962) attribute the drop in recognition rate introduced by the smoothing of the contours to the presence of micro-perturbations ("jitter") in some expressions. Those micro-perturbations of the F0 contours allowed to differentiate some expressions of joy and fear, which were mistaken only when the resynthesized contours were smoothed. In this study, more long term variations of the F0 contours—comparable to those that would be captured by a linguistic transcription model—also allowed to differentiate the "emotional modes" portrayed but to a less important extent (only one expression out of four is still correctly categorized by the listeners in this condition).

In a series of three studies using resynthesized speech, Ladd et al. (1985) assessed the effect of a combined manipulation of F0 contour shape ("uptrend" versus "downtrend") and stepwise increase of F0 range on emotional attributions. They found that contour shape, F0 range and also voice quality (one speaker produced expressions using two different phonation modes) independently influenced emotional ratings. They also reported that the progressive increase of F0 range affected emotional intensity ratings to a greater extent than the manipulation of the contour shape.

Uldall (1964) applied 16 stylized F0 contours on five utterances. She showed that the emotional meaning attributed to different contours varies depending on the sentence carrying the contour. She found for instance that the contour featuring a weak declination and a low level is rated as 'unpleasant', 'authoritative', and corresponding to a 'weak' emotional intensity when applied to the two types of questions and the statement used in this study. The same contour is rated as 'unpleasant', 'authoritative' and corresponding to a 'strong' emotional intensity when applied to the command utterance. Uldall identified only few contour features that were linked to the three dimensions (valence, power, intensity) underlying the emotional ratings of the participants in this study, independently of the three sentence types carrying the contours.

Likewise, Scherer et al. (1984) found that specific combinations of prosodic features (final rise or fall of F0 contours) and linguistic categories (Wh-questions or yes/no questions) influence the attributions of affect-loaded attitudes (such as "challenging", "agreeable" or "polite"). A final fall of the F0 contour will for example be perceived as "challenging" on a yes/non question (where a final rise is expected form the syntactic structure)

[2] The eight categories selected in this study were closer to interpersonal attitudes than to full-blown emotions: (1) 'bored statement', (2) 'confidential communication', (3) 'question expressing disbelief', (4) 'message expressing fear', (5) 'message expressing happiness', (6) 'objective question', (7) 'objective statement', (8) 'pompous statement'.

but not on a Wh-question. In this study as well, perceived emotional intensity was affected mainly by the continuous variation of F0 level. The authors suggested that vocal aspects covarying with emotional attributions (such as F0 level in this study) might mainly reflect and communicate the physiological arousal associated to the emotional reaction, whereas configurations of prosodic features (such as F0 contour shapes) would be used to signal specific attitudes in association with the linguistic content of the utterances.

On the whole, the literature does not provide strong evidence for the existence of emotion specific intonation patterns. Nevertheless, intonation (or more specifically F0 fluctuations) seem to be affected to some extent by the emotional state of the speakers and appear to carry information that can be used by listeners to generate inferences about the emotional state of the speakers, more or less independently of the linguistic features of the expressions. The notion of emotional intonation being produced and processed independently of the linguistic aspects of speech is further supported by neuropsychological studies conducted to differentiate structures involved in the processing of linguistic intonation, on one hand, and emotional intonation, on the other hand (e.g. Heilman et al., 1984; Pell, 1998; Ross, 1981; van Lancker and Sidtis, 1992).[3] Furthermore, studies investigating the prelinguistic production and perception of intonation in infants suggest that emotional meaning can be communicate through modulations of intonation before language is acquired (Fernald, 1991, 1992, 1993; Papousek et al., 1991).

Altogether then, despite the relative lack of empirical evidence, there are strong claims supporting the notion that F0 contours can carry emotional meaning independently of linguistic structures. The study we present in this paper introduces a new, more resolutely quantitative approach to the description of F0 fluctuations in emotional speech including a more elaborate description of F0 contours than the aggregated F0 descriptors mostly used in empirical studies on vocal correlates of emotional expressions.

[3] Although, the possibility to clearly separate linguistic and emotional intonation is also debated in this field (e.g. Snow, 2000).

2. Empirical assessment of F0 contours in emotion portrayals

2.1. Purpose

The purpose of the study was to examine whether we could find specific contour types for a number of emotions, using a large corpus of vocal emotion expressions that has been extensively studied for voice quality, aggregated intensity and temporal measures, and aggregated F0 measures (Banse and Scherer, 1996; Bänziger, 2004). The first task was to develop a contour coding system that is amenable to quantitative statistical analysis.

2.2. Development of an appropriate intonation coding system

In our opinion, linguistic models of intonation are not the most appropriate for the description and analysis of pitch in emotional expressions. First, we argue that the distinctive categories (tones or contours) used to describe linguistic pitch variations are not well suited to describe variations of pitch involved in emotional communication. Results of past studies (Ladd et al., 1985; Scherer et al., 1984) suggest, for instance, that the effects of emotions on intonation are likely to be continuous rather than categorical. Furthermore, quantitative descriptions of pitch contours would allow us not only to account for continuous variations of pitch dimensions, but also to statistically analyze and compare pitch dimensions for various emotional expressions. Finally, linguistic models of intonation are concerned mainly with perceived fluctuations of pitch. The reliance on perception to describe and analyze intonation entails several problems: (a) Perceived pitch fluctuations are influenced by interactions of multiple factors on the level of speech production (e.g., F0, intensity, duration, spectral distribution of energy). Transcriptions of perceived pitch are therefore not very informative regarding the aspects of voice

production and voice signals that are affected by portrayed emotions. (b) Evaluations of perceived pitch are highly subjective; they could be biased by expectancies of the coders and/or influenced by the emotions perceived in the vocal expressions. (c) The subjectivity of the coding of perceived pitch is likely to lead to low inter-coder reliabilities (on this issue, see Syrdal and McGory, 2000; Wightman, 2002).

Consequently, we favor a quantitative approach of pitch contour description and analysis, oriented toward the voice signal (i.e., remote from the perceived categorization of pitch fluctuations). Furthermore, in the corpus we describe in the following paragraphs, meaningless sequences of syllables were produced by actors who portrayed a comprehensive set of emotions. Identification of speech segments relying, implicitly or explicitly, on syntactic or semantic aspects was therefore impossible, and linguistic models of intonation are hardly applicable. Therefore, we decided to develop a stylization/coding procedure for F0 contours for our own purposes that requires a minimal set of assumptions about the underlying phonetic and syntactic structure. This stylization procedure was inspired by the work of Patterson and Ladd (1999) on pitch range modeling. A set of objective criteria were defined and used for the stylization of F0 contours in order to reduce the influence of the subjective interpretation of the coder on the description of the pitch contours. An external speaker baseline was introduced in order to compare features of F0 contours for different emotional expressions. The corpus of emotional expressions and the procedure used for the stylization of the F0 contours are described as follows.

2.3. Characteristics of the emotional expressions used in this study

2.3.1. Portrayed emotions

The corpus used in this study consisted of 144 emotional expressions, which were sampled from a larger set of 1'344 emotional expressions described in detail by Banse and Scherer (1996). Expressions produced by nine professional actors (four males and five females) were selected. All actors pronounced two sequences of seven syllables (1. "hät san dig prong nju ven tsi." 2. "fi gött laich jean kill gos terr"). Four emotion categories (or families)—anger, fear, sadness and joy—were crossed with two levels of emotional arousal (or activation)—low arousal ('LA') versus high arousal ('HA')—resulting in eight emotions: cold anger ('LA anger') and hot anger ('HA anger'), anxious fear ('LA fear') and panic fear ('HA fear'), depressed sadness ('LA sad') and despaired sadness ('HA sad'), calm joy/happiness ('LA joy') and elated joy ('HA joy'). This combination between emotion category and arousal level is an essential feature of this data set. Emotional arousal is known to substantially influence vocal expressions. Emotions involving high arousal have been consistently described as being expressed with louder voice, faster speech rate, higher pitch (etc.) than emotions with low arousal. Whereas, in most past studies, emotions involving very low arousal such—as depressed sadness or boredom—have been directly compared with emotions involving very high arousal such as panic fear or hot anger, the crossing of arousal and emotion category in this data set allows to separate their respective influence on vocal expressions, and to assess the possibility to differentiate emotions with similar arousal level.

All actors were native German speakers and were provided with short scenarios for each emotion to ensure that they share the same definitions for the labels provided.[4] Hence, the emotions portrayed in the data set are defined by the expressive intentions of the actors based on those labels and scenarios. The advantages and drawbacks of using acted speech to study emotional expressions have been extensively discussed elsewhere (e.g. Banse and Scherer, 1996). The assumption is made here that although acted emotional expressions may be deprived of certain characteristics pertaining to spontaneous expressions—such as subtle

[4] The german labels provided to the actors were the following: *Angst* (LA fear), *Panische Furcht* (HA fear), *Stille Trauer* (LA sad), *Verzweiflung* (HA sad), *Stille Freude* (LA joy), *Überschäumende Freude* (HA joy), *Kalter Ärger* (LA anger), *Heisser Ärger* (HA anger). The scenarios can be obtained upon request to the authors. More details regarding the construction of the scenarios can be found in Banse and Scherer (1996).

modifications of voice quality related to physiological modifications that actors might not succeed in producing deliberately—and may in some instances sound "theatrical", they still represent close approximations of genuine emotional expressions. The emotional communication potential of the portrayals was assessed in a recognition study. The results of this study are briefly described in Section 2.3.4.

One instance of each portrayed emotion (8) was selected randomly for each speaker (9) and each sequence of syllables (2) among the emotional expressions featured in the broader data set, yielding a total of 144 ($8 \times 9 \times 2$) emotional expressions examined in this study.[5]

2.3.2. Verbal content

The actors pronounced two sequences of syllables: (1) "hät san dig prong nju ven tsi.", (2) "fi gött laich jean kill gos terr". Those meaningless "sentences" combine phonetic elements from various European languages and were constructed to sound like an unknown language for lay listeners with different linguistic backgrounds (Scherer et al., 1991). Compared to real sentences, the linguistic constraints on the pitch contours are obviously reduced in those utterances, which are deprived of a given syntactic and semantic structure. Nevertheless, the speakers (German actors) might have allocated syntactic structure(s) and/or meaning(s) to the utterances while producing the expressions. French-speaking participants in the recognition study (Section 2.3.4) often declared that they thought this "language" to be an unknown Germanic language. The pronunciation of some of the phonemes included in the utterances might account for this impression, but it cannot be ruled out that the intonation of the expressions produced by the German speakers also showed recognizable Germanic specificities. Concisely, in the absence of a designated syntactic structure or meaning, the actors were free to choose the intonation they preferred and to vary it for the sake of emotional communication, but they most likely applied the rules of their native language to produce the intonation of the recorded utterances.[6]

2.3.3. Aggregated F0 measures

F0 was extracted by autocorrelation using the speech analysis software PRAAT (Boersma and Weenink, 1996). Detection errors—octave jumps and detection of periodicity in unvoiced speech segments—were manually corrected; but no correction was applied when the algorithm failed to detect periodicity in voiced segments. Aggregated F0 descriptors—such as F0 mean, F0 minimum and F0 range—were computed for each expression. Speaker variance was subtracted by standardizing the extracted parameters for each speaker. Fig. 1 provides an overview of the profiles obtained for aggregated F0 measures. Average values and standard deviations are represented for the eight emotions included in the data set. Emotions including low levels of arousal are represented on the left side, whereas emotions including high levels of arousal are represented on the right side in this figure. Aggregated F0 measures tend to be higher for emotions with high arousal than for emotions with low arousal, but some exceptions to this overall effect can be observed as well. F0 range is for instance relatively high (on average) for 'cold anger' (LA anger), reaching the level of the average F0 range of highly aroused expressions. Furthermore, differences for emotions with similar arousal can be observed already at this broad level of F0 description; F0 minimum is for instance higher on average for 'panic fear' (HA fear) than for 'hot anger' (HA anger).

2.3.4. Emotional attributions

In this data set, the emotional labels assigned to the vocal expressions are defined by the expressive intentions of the actors, who were given emotion

[5] The authors will provide the 144 recordings selected for this study upon written request and based on agreement regarding their potential utilization.

[6] Consequently, we do not claim to look for *universal* pitch contours in this study. We only considered pitch contours produced by German speakers in the absence of a given syntactic and semantic structure (i.e. where the intonation cannot be derived from the words, the syntax or the meaning). A case for *universality* could be made only if native speakers of different languages would produce the same pitch contours for the same emotions (on different or same linguistic contents).

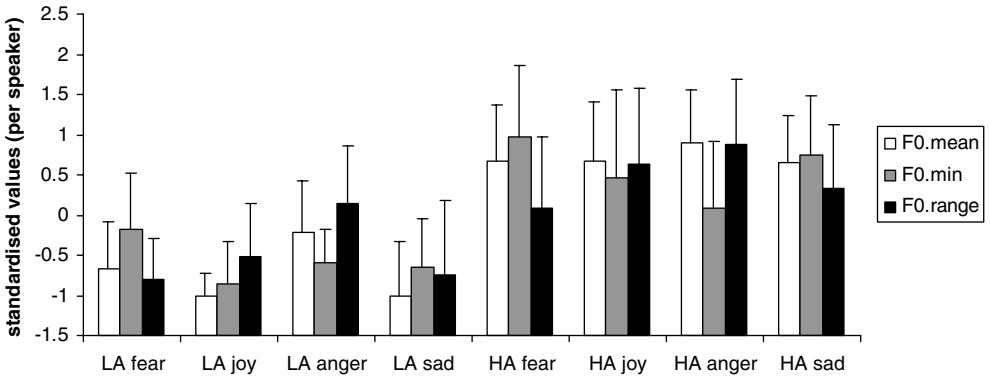

Fig. 1. Aggregated measures of F0, mean and standard deviation by portrayed emotion. *Note*: F0 was measured in Hz. The three F0 descriptors (F0 mean, F0 min and F0 range) were extracted for each expression. The obtained values were then standardized separately for each speaker to remove the speaker variance (i.e. a linear transformation was applied to the obtained values for each speaker). The resulting values (vertical scale on Fig. 1) average to 0 for all expressions.

words and scenarios to characterize the categories they should express. Accordingly, the objective of this study is to describe fluctuations of F0 associated with expressive intentions. Nevertheless, the emotions attributed to the expressions produced by the actors are also of interest, in so far as they provide information about the communication potential of the expressions under investigation.

Ratings of emotional intensity were obtained from French-speaking listeners (first year psychology students at the University of Geneva). While the recordings to be assessed are usually displayed in fixed or random order and are evaluated by listeners on a number of different scales immediately after the display, we used a different approach in this study. Four visual-analogue rating scales were presented successively to the listeners on a computer screen in random order. The scales represent the perceived intensity of joy, the perceived intensity of anger, the perceived intensity of fear and the perceived intensity of sadness, ranging from no 'emotion' (joy, anger, fear, or sadness) to extreme 'emotion' (joy, anger, fear, or sadness). The task of the listeners is to position the recordings on each successive scale. All recordings produced by one given speaker/actor (16 expressions) are represented simultaneously on the screen as identical icons; the participants display (listen to) the recordings by double-clicking the icons and move them to the position they select on the visual-analogue scale. The participants can listen to the recordings and modify/correct their answers as often as they need. The answers are recorded on a continuous scale ranging from 0 (for recordings positioned on the extreme left of the scale, no emotion) to 10 (for recordings positioned on the extreme right of the scale, extreme emotion). The direct comparison of all the expressions produced by a speaker allows the listeners to adjust their ratings to the range of the expressions produced by the speaker and prevents shifts in internal comparison standards over time. Four separate groups of listeners rated the total set of 144 expressions. Each group assessed the emotional expressions produced by three different speakers. In total, ratings from 14 participants were obtained for each emotional expression.

Fig. 2 displays average ratings of perceived intensity of joy, perceived intensity of fear, perceived intensity of anger, and perceived intensity of sadness for the eight emotions included in the data set. The error bars represent the variability (standard deviation) of the ratings for different expressions belonging to the same emotion category. The principal effect of arousal level on emotional intensity attributions can be described as follows: expressions of 'elated joy' (HA joy), 'panic fear' (HA fear) and 'hot anger' (HA anger)—i.e. expressions involving high activation—trigger attributions of, respectively, more intense joy,

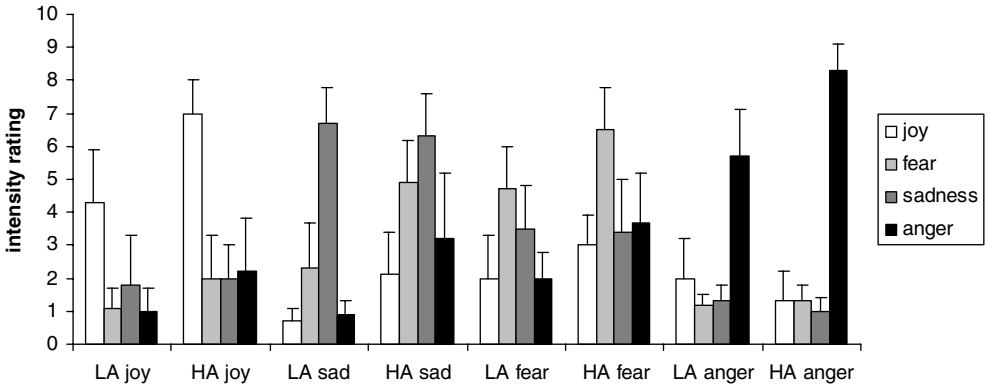

Fig. 2. Average ratings of emotional intensity (perceived intensity of joy, of fear, of sadness, and of anger) by portrayed emotion.

more intense fear, and more intense anger than expressions of 'calm joy' (LA joy), 'anxious fear' (LA fear), and 'cold anger' (LA anger)—i.e. expressions of corresponding emotion categories but involving low activation; unlike expressions of depressed sadness (LA sad), which were rated as being as sad as expressions of despaired sadness (HA sad). Confusions between portrayed and perceived emotion categories occurred only for expressions of 'despair' (HA sad), which received ratings of fear intensity as high as for expressions of 'anxious fear' (LA fear).

The data presented above indicate that listeners are able to discriminate the expressions used in this study in a way that can be, at least partially, predicted from the expressive intentions of the actors. They also show that arousal level is not directly predictive of perceived emotional intensity. In our data set, the relation between emotional arousal and perceived emotional intensity varies for different categories of portrayed emotions. The results also indicate that listeners are probably able to discriminate emotion categories even when they include "unusual" levels of activation.[7] Furthermore, descriptions of the aggregated F0 measures provided in Fig. 1 show that F0 is affected by the emotion categories included in our data set and that arousal level only cannot account totally for this effect. The following paragraphs describe the procedure used for the coding of the F0 contours.

2.4. Procedure used for the stylization

Ten key points were identified for each F0 contour. The first point ('start') corresponds to the first F0 point detected for the first voiced section in each expression. This point is measured on the syllable "hät" in sequence 1 and on the syllable "fi" in sequence 2. The second ('1min1'), third ('1max'), and fourth points ('1min2') correspond, respectively, to the minimum, maximum, and minimum of the F0 excursion for the first operationally defined "accent" of each sequence. Those local minima and maxima are measured for the syllables "san dig" in sequence 1 and for the syllables "gött laich" in sequence 2. Point five ('2min1'), six ('2max'), and seven ('2min2') correspond, respectively, to the minimum, maximum, minimum of the F0 excursion for the second operationally defined "accent" of each sequence. They are measured for the syllables "prong nju ven" and "jean kill gos," Point eight ('3min'), nine ('3max'), and ten ('final') correspond to the final "accent" of each sequence: the local minimum, maximum, and minimum for the syllables "tsi" and "ter." Fig. 3a shows an illustration of this stylization for a happy (LA joy) expression, while Fig. 3b shows another illustration of the stylization for an expression of hot anger (HA anger); both expressions are produced on utterance 1. The

[7] In most studies of vocal expressions, anger is displayed with high arousal and sadness with low arousal only. Therefore, there is very little evidence concerning the possibility for listeners to correctly identify sad expressions when they include high arousal or angry expressions when they include low arousal.

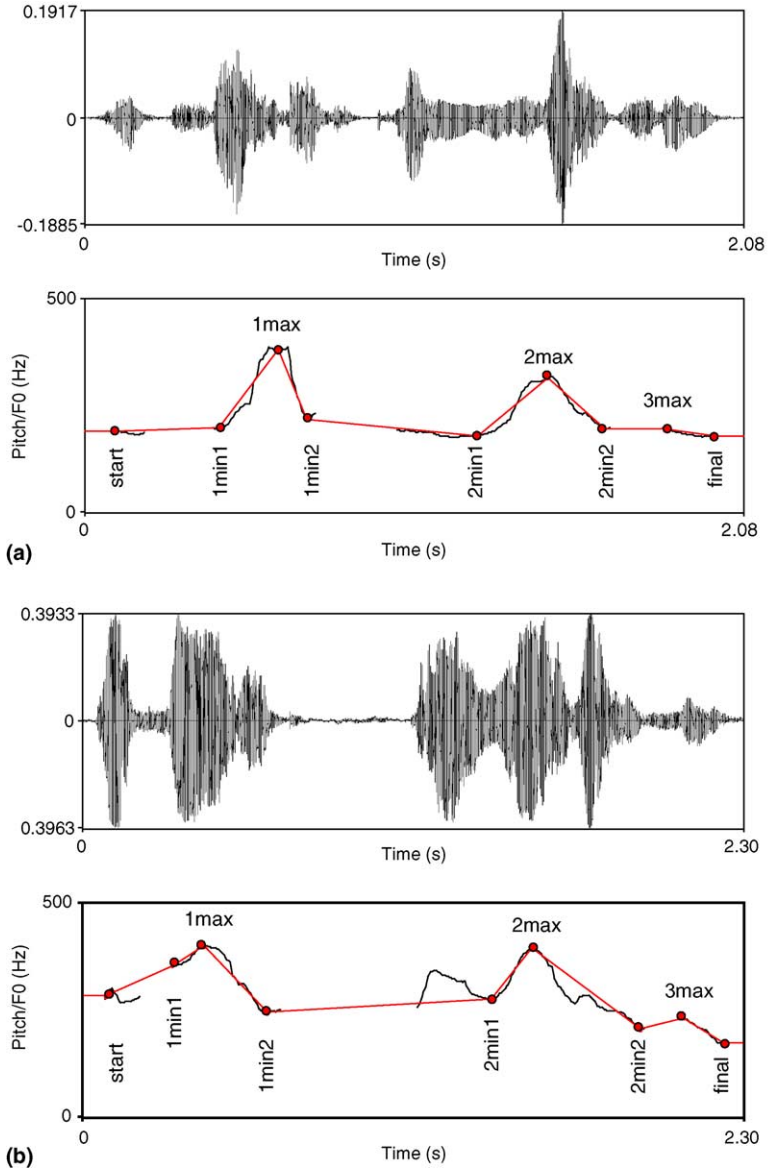

Fig. 3. (a) Stylization example (an instance of LA joy (a) and HA anger (b) with seq 1).

original F0 contours are represented by black curves; the stylized contours are superimposed in straight lines (larger dots represent the manually/visually identified local minima/maxima). Point eight ('3min') is missing in both expressions. F0 fluctuations that did not correspond to the criteria described earlier were ignored. An example is presented in Fig. 3b. On the 4th syllable ("prong"), the F0 excursion was ignored; only one excursion (on the 5th and 6th syllables) is coded for the 2nd group of syllables "prong nju ven."

2.5. Results

The pattern represented in Fig. 3—two "accents" (sequences of local F0 min1-max-min2)

followed by a final fall—was the most frequent pattern for the 144 expressions submitted to this analysis. The count of F0 ''rises'' (local 'min1' followed by 'max'), ''falls'' (local 'max' followed by 'min2'), and ''accents'' ('min1' followed by 'max' followed by 'min2') for the first accented part, the second accented part, and the final syllable, as shown in Table 1, was not strongly affected by the portrayed emotions, but varied much more over different speakers and the two sequences of syllables that they pronounced (e.g., there were only 5 occurrences of the point '3min' for sequence 1, versus 43 occurrences of this point for sequence 2). The amount of observations in single cells is too low to allow to perform a statistical analysis on this data.

In order to control for differences in F0 level between speakers, a ''baseline'' value was defined for each speaker. An average F0 value was computed on the basis of 112 emotional expressions (including the 16 expressions used in this study) produced by each speaker. Fig. 4 shows the differences in Hz (averaged across speakers and sequences of syllables) between the observed F0 points in each expression and the speaker baseline value for each portrayed emotion.

Fig. 4 shows that F0 level is mainly affected by emotional arousal. The F0 points for emotions with low arousal—such as 'sadness' (LA sad), 'calm joy' (LA joy), and 'anxious fear' (LA fear)—are generally lower than the F0 points for emotions with high arousal—'despaired sadness' (HA sad), 'elated joy' (HA joy), 'panic fear' (HA fear), and 'hot anger' (HA anger). The general F0 level of recordings expressing 'cold anger' (LA anger) is located between the F0 level of other low aroused and of high aroused expressions.

The description of the different points in the contour does not appear to add much information to an overall measure of F0, such as F0 mean. Looking at the residual variance after regressing F0 mean (computed for each expression and subtracted with the speaker baseline) on the points represented in Fig. 5, a series of ANOVAs computed separately on the nine measurement points reveal that there remains only a slight effect of portrayed emotion on point '2max': $F(7,125) = 6.91$, $p < 0.001$, $\eta^2 = 0.28$ and point 'final': $F(7,102) = 4.42$, $p < 0.001$, $\eta^2 = 0.23$. Posthoc tests (Tukey's HDS, $p < 0.05$) show that the second maximum tends to be higher for recordings expressing 'elated joy' (HA joy), 'hot anger' (HA anger), and 'cold anger' (LA anger) than for recordings expressing depressed 'sadness' (LA sad) and 'anxious fear' (LA fear). The second maximum is also higher for 'elated joy' (HA joy) and 'hot anger' (HA anger) than for 'calm joy' (LA joy). The final F0 value tends to be relatively lower for 'hot anger' (HA anger) than for 'despaired sadness' (HA sad), 'panic fear' (HA fear), 'depressed sadness'

Table 1
Count of observed F0 points (rise, fall, accent) by portrayed emotion and by sequence of syllables

Segment	F0 excursion	LA fear	HA fear	LA anger	HA anger	LA joy	HA joy	LA sad	HA sad	seq 1	seq 2
Start	Observed	14	16	18	18	16	18	14	16	64	66
	Absent	4	2	0	0	2	0	4	2	8	6
acc1	Accent	10	9	8	12	11	12	9	11	50	32
	Rise	5	8	8	4	4	5	5	1	12	28
	Fall	1	1	2	0	1	0	3	2	4	6
	Absent	2	0	0	2	2	1	1	4	6	6
acc2	Accent	16	16	16	15	16	15	11	13	64	54
	Rise	0	1	0	1	0	1	1	1	1	4
	Fall	1	0	1	0	0	1	5	2	5	5
	Absent	1	1	1	2	2	1	1	2	2	9
Final	Accent	4	6	5	4	5	9	1	6	3	37
	Rise	0	1	0	1	2	1	3	0	2	6
	Fall	8	9	11	11	9	6	6	10	49	21
	Absent	6	2	2	2	2	2	8	2	18	8

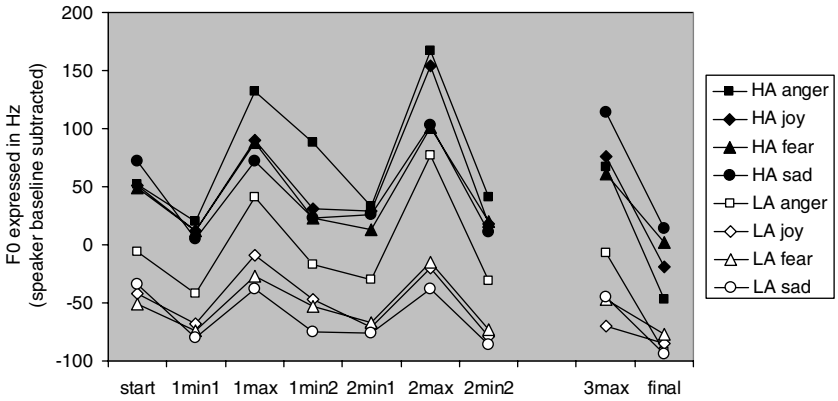

Fig. 4. Average F0 values by portrayed emotion. *Note*: The number of observations varies from 18 (for 'start' with hot anger, cold anger and elation; for '1max' with cold anger and panic fear) to 7 (for 'final' with sadness). It should be noted also that there is a sizeable amount of variance around the average values shown for all measurement points.

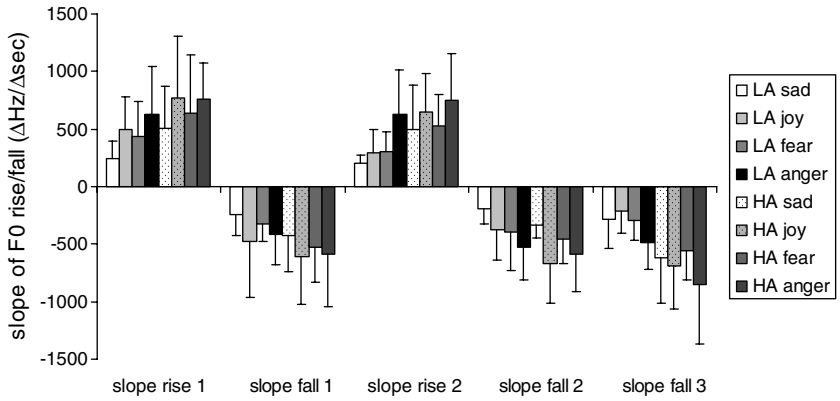

Fig. 5. Rising and falling F0 slopes, means, and standard deviations per portrayed emotion.

(LA sad), 'calm joy' (LA joy), and 'anxious fear' (LA fear).

Slopes for rising segments of the stylized F0 were computed by subtracting the first local minimum (point '1min1' or '2min1' in Hz) from the local maximum ('1max' or '2 max', respectively, in Hz) and then dividing this difference by the duration (in seconds) of the F0 excursion between the first local minimum and the local maximum. Slopes for falling segments of the stylized F0 were computed by subtracting the local maximum (point '1max,' '2max,' or '3 max,' in Hz) from the second local minimum (respectively, '1min2,' '2min2,' or 'final' in Hz) and then dividing this difference by the duration (in seconds) of the F0 excursion between the local maximum and the second local minimum.

The average values (and standard deviations) of the rising and falling slopes for each portrayed emotion are presented in Fig. 5. The slopes tend to be steeper for part of the high-aroused emotions—especially for 'elation' (HA joy) and 'hot anger' (HA anger)—and less steep for part of the low-aroused emotions—especially for 'sadness' (LA sad), 'joy' (LA joy), and 'anxiety' (LA fear). The similarity of the patterns observed on the five slopes for different emotions suggests that a more global evaluation of F0 range might account for the differences between emotions on all slopes. To test this assumption, we regressed F0 range—defined

as the difference between the absolute minimum and the absolute maximum in each expression—on the five slopes. The effect of the portrayed emotions on the residuals of this regression was assessed by a series of five ANOVAs. After we controlled for the influence of F0 range, emotions did not affect the slopes of the first F0 excursion any longer. Small differences remained essentially for the second F0 rise—$F(7,114) = 2.07$, $p = 0.052$, $\eta^2 = 0.11$—with, for instance, steeper slopes for 'elation' (HA joy), 'cold anger' (LA anger), and 'hot anger' (HA anger) than for 'sadness' (LA sad) and 'happiness' (LA joy), and for the final fall—$F(7,100) = 2.73$, $p = .012$, $\eta^2 = 0.16$—with, for instance, a steeper fall for 'hot anger' (HA anger) than for the low-aroused emotions (LA joy, LA fear, LA sad, LA anger) and for 'panic fear' (HA fear).

Additionally, the relative location of the absolute maximum of F0 (F0 peak) in the expressions was examined. The most important observation in this respect was a remarkable difference between the average location of the maximum F0 for happy expressions (LA joy) and the average location of the maximum F0 for elated expressions (HA joy). For most happy expressions, F0 peak was reached on the second segment of the expressions (''san dig''/''gött laich''), whereas for most elated expressions, F0 peak was reached on the third or final segments (''prong nju ven—tsi''/''jean kill gos—terr'') of the expressions. On average, F0 peak was measured at 46% of the total duration of the utterances for happy expressions, and at 72% of the utterances for elated expressions.

Furthermore, the second local maximum was significantly higher than the first local maximum in expressions of 'cold anger' (LA anger), 'panic fear' (HA fear), 'despaired sadness' (HA sad), and 'elated joy' (HA joy). In addition there was a significant decrease from the second local maximum to the third local maximum in all emotional expressions except for expressions of 'despaired sadness' (HA sad) and 'elated joy' (HA joy). In other words, the ''accentuation'' of despaired and elated expressions is, on average, more marked on the second part than on the first part of the utterances, and the ''F0 ceiling'' is consistently higher in those expressions until the final fall than in expressions of 'cold anger' (LA anger) and 'panic fear' (HA fear), which are also more ''accentuated'' on the second part than on the first part of the utterances. Expressions of 'sadness' (LA sad), 'happiness' (LA joy), 'anxiety' (LA fear), and 'hot anger' (HA anger) did not show more ''accentuation'' on the second part than on the first part of the utterance. In addition, the ''F0 ceiling'' of those expressions is notably lowered before the final fall on the last syllable.

Finally, the global ''declination''—defined as the difference (in Hz) between the first measured value of F0 on the first syllable (''hät''/''fi'') and the last measured value of F0 on the final syllable (tsi/terr), divided by the duration (in seconds) separating those two points—was examined. Fig. 6 shows the means and standard deviations of this ''declination'' for each portrayed emotion. The variance within each emotion being very important (see standard deviations in Fig. 6) and the

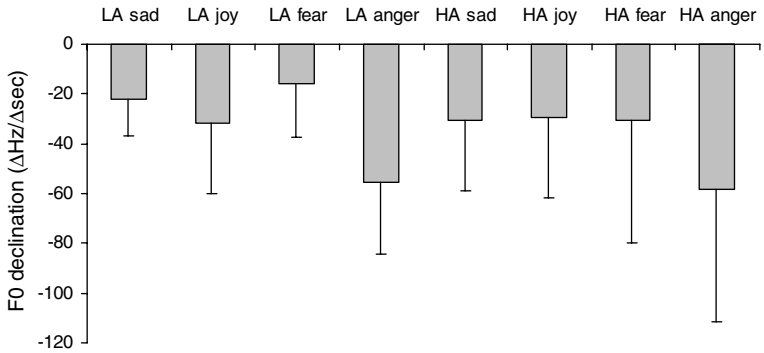

Fig. 6. Mean and standard deviation of F0 declination per portrayed emotion.

number of expressions analyzed being relatively small, the differences between portrayed emotions are not significant. Statistically, the F0 declination of expressions corresponding to 'hot anger' (HA anger—58 Hz per second) only tends to be steeper than the F0 declination of expressions corresponding to 'anxiety' (LA fear—16 Hz per second).

3. Conclusions

The results of our quantitative prosodic analysis of a large corpus of vocal emotion portrayals indicate that the level of arousal underlying portrayed emotions essentially affected the global level and range of F0 contours. Therefore, simple summaries of F0 contours—such as F0 mean or F0 range—were sufficient to account for the most important variations observed between emotion categories.

However, a more detailed examination of the contours revealed specific differences for some portrayed emotions. For some emotional expressions—especially hot anger (HA anger), cold anger (LA anger), and elation (HA joy)—the second F0 excursion in the utterances tended to be larger than for other emotions—such as sadness (LA sad) or happiness (LA joy), which showed much smaller F0 excursions in the second part of the utterances. This difference could not be explained entirely by the overall difference in F0 range for those expressions.

The "shape" of the contours was only slightly affected by the portrayed emotions. Contours with "uptrend" shape (a term borrowed from Ladd et al., 1985)—i.e., contours featuring a progressive increase of F0 and upholding a high level of F0 until the final fall—were observed for expressions of despair (HA sad) and elation (HA joy), whereas expressions of sadness (LA sad) and happiness (LA joy) showed a "downtrend" movement of F0—an early F0 peak followed by a progressive decrease until the final fall. The final fall itself might also be affected by portrayed emotions. Emotions such as hot anger (HA anger) or elation (HA joy) might result in steeper final falls than expressions of anxiety (LA fear) or happiness (LA joy).

The results regarding the relative height of local F0 excursions, contour "shape," and final fall must be considered with caution. The variations within portrayed emotions were always large in the corpus we examined and the number of expressions analyzed was relatively small. Consequently, those results need to be replicated before they can be generalized.

The coding of F0 contours used in this study might have cancelled out potentially important aspects of the F0 contours. Specific configurations of F0 contour features and syntactic aspects (as described in (Scherer et al., 1984); or (Uldall, 1964)), semantic aspects, or even phonetic aspects of the expressions might contribute to the expression and the communication of emotional meaning. The expressions considered in this study were free of semantic and syntactic content that might have allowed to communicate an emotional impression in interaction with F0 contour features. Still the local F0 excursions ("accents") observed on those expressions were produced with large variations regarding their precise location in the utterances. These variations were cancelled out by the coding of the F0 contours. Therefore, the possibility remains that a more thorough examination of the position of the F0 excursions relatively to the phonetic content of the utterances might allow to identify emotion specific differences.

On the other hand, it should be noted that, in the absence of syntactic or semantic constraints, the actors were free to choose the contour that would have seemed best suited to convey a particular emotional feeling. The fact that they did not systematically produce such emotion-specific contours for those short utterances seems to indicate that emotions considered independently of linguistic context do not provide for contour coding other than the general level, range, and final fall parameters described earlier.

As mentioned earlier, these results certainly need to be replicated and it would probably be useful to include a number of utterances that do have linguistic structure and meaning to compare with the kinds of quasi-speech stimuli that we have been using. It would also be beneficial to systematically record portrayals of affect bursts (Schröder, 2003). As mentioned earlier, one would need to agree on

an intonation coding system that respects both the needs of statistical analysis and fundamental aspects of contour shape, without getting into the subtleties of the debates between schools in linguistics and phonology. Obviously, it would be useful if such a system worked mostly automatically, with hand correction. Once we have the appropriate corpus, preferably produced with actors from different cultures and language groups, we could use some of the techniques for signal masking and feature destruction that allow us to determine which aspects of a signal need to be retained to carry recognizability. The fact that, in the past, random-splicing procedures (which destroy intonation and sequential information but keep voice quality) have worked better, in the sense of preserving recognition accuracy, than content-filtering methods (which keep intonation but mask essential aspects of voice quality; Scherer et al., 1985) suggests that intonation contours (at least in terms of shape) may be less important signatures of emotions than global F0 level and variation and spectral aspects of voice quality.

Finally, emotional speech synthesis should be the method of choice to systematically test the hypotheses that have been obtained by the more exploratory methods. Although the commercial interest in affect-rich multimodal interfaces has led to a multiplication of emotion synthesis studies, to our knowledge few have advanced to a satisfactory level of ecological validity. All too often, such work either is not based on hypotheses informed by earlier work, or suffers from serious methodological shortcomings (e.g., inflated recognition rates due to a limited number of categories and failure to distinguish simple discrimination from pattern recognition). One of the major problems is that engineers and phoneticians, but unfortunately also some psychologists, tend to think that emotions are easy to understand and to manipulate and that we understand them because we experience them ourselves. Nothing could be further from the truth. The vocal expression of emotion may be one of the most complex systems of communication there is, certainly much more complex than facial expression. In consequence, advances in the field should rely, much more than in the past, on close collaboration between phoneticians, speech scientists, engineers, and psychologists.

References

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636.

Bänziger, T., 2004. Communication vocale des émotions: perception de l'expression vocale et attributions émotionnelles. Unpublished doctoral thesis.

Boersma, P., Weenink, D.J.M., 1996. Praat, a System for Doing Phonetics by Computer, Version 3.4 (132). Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam.

Cruttenden, A., 1986. Intonation. Cambridge University Press, Cambridge.

Fernald, A., 1991. Prosody in speech to children: Prelinguistic and linguistic functions. In: Vasta, R. (Ed.), Annals of Child Development. Jessica Kingsley Publishers, London, pp. 43–80.

Fernald, A., 1992. Meaningful melodies in mothers' speech to infants. In: Papousek, H., Juergens, U. (Eds.), Nonverbal Vocal Communication: Comparative and Developmental Approaches. Cambridge University Press, Cambridge, pp. 262–282.

Fernald, A., 1993. Approval and disapproval: Infant responsiveness to vocal affect in familiar and unfamiliar languages. Child Development 64, 657–674.

Fonagy, I., Magdics, K., 1963. Emotional patterns in intonation and music. Zeitschrift für Phonetik 16, 293–326.

Frick, R.W., 1985. Communicating emotion: The role of prosodic features. Psychological Bulletin 97, 412–429.

Fujisaki, H., 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: Fujimura, O. (Ed.), Vocal Physiology: Voice Production, Mechanisms and Functions. Raven, New York, pp. 347–355.

Halliday, M.A.K., 1970. A Course in Spoken English: Intonation. Oxford University Press, Oxford.

Heilman, K.M., Bowers, D., Speedie, L., Coslett, H.B., 1984. Comprehension of affective and nonaffective prosody. Neurology 34, 917–921.

Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin 129, 770–814.

Ladd, D.R., Silverman, K., Tolkmitt, F., Bergmann, G., Scherer, K.R., 1985. Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. Journal of the Acoustical Society of America 78, 435–444.

Léon, P.R., Martin, P., 1970. Prolégomènes à L'étude Des Structures Intonatives. Marcel Didier, Montréal.

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

Mozziconacci, S.J., 1998. Speech variability and emotion: production and perception. Doctoral thesis. Technische Universiteit Eindhoven, Eindhoven.

O'Connor, J.D., Arnold, G.F., 1973. Intonation of Colloquial English, Second ed. Longman, London.

Pakosz, M., 1983. Attitudinal judgments in intonation: Some evidence for a theory. Journal of Psycholinguistic Research 12, 311–326.

Papousek, M., Papousek, H., Symmes, D., 1991. The meanings of melodies in motherese in tone and stress languages. Infant Behavior and Development 14, 415–440.

Patterson, D., & Ladd, D.R., 1999. Pitch range modelling: linguistic dimensions of variation. Paper presented at the XIVth International Congress of Phonetic Sciences (ICPhS), San Francisco.

Pell, M.D., 1998. Recognition of prosody following unilateral brain lesion: Influence of functional and structural attributes of prosodic contours. Neuropsychologia 36, 701–715.

Pierrehumbert, J., 1980. The Phonology and Phonetics of English Intonation. Massachusetts Institute of Technology.

Ross, E.D., 1981. The aprosodias: Functional-anatomic organization of the affective components of language in the right hemisphere. Annals of Neurology 38, 561–589.

Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256.

Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T., 1991. Vocal cues in emotion encoding and decoding. Motivation and Emotion 15, 123–148.

Scherer, K.R., Feldstein, S., Bond, R.N., Rosenthal, R., 1985. Vocal cues to deception: A comparative channel approach. Journal of Psycholinguistic Research 14, 409–425.

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, 1346–1356.

Schröder, M., 2003. Experimental study of affect bursts. Speech Communication. Special Issue Speech and Emotion 40, 99–116. Accessible at: http://www.dfki.de/~schroed.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P., Hirschberg, J., 1992. ToBI: A standard scheme for labeling prosody. Paper presented at the International Conference on Spoken Language Processing (ICSLP), Beijing, China.

Snow, D., 2000. The emotional basis of linguistic and nonlinguistic intonation: Implications for hemispheric specialization. Developmental Neuropsychology 17, 1–28.

Syrdal, A.K., McGory, J., 2000. Inter-transcriber reliability of ToBI prosodic labeling. Paper presented at the International Conference on Spoken Language Processing (ICSLP), Beijing, China.

t'Hart, J., Collier, R., Cohen, A., 1990. A Perceptual Study of Intonation. Cambridge University Press, Cambridge.

Uldall, E., 1964. Dimensions of meaning in intonation. In: Abercrombie, D., Fry, D.B., MacCarthy, P.A.D., Scott, N.C., Trim, J.L.M. (Eds.), In Honour of Daniel Jones: Papers Contributed on the Occasion of His Eighteenth Birthday, 12 September 1961. Longman, London, pp. 271–279.

van Lancker, D., Sidtis, J.J., 1992. The identification of affective-prosodic stimuli by left- and right-hemisphere-damaged subjects: All errors are not created equal. Journal of Speech and Hearing Research 35, 963–970.

Wightman, C.W., 2002. ToBI or not ToBI. Paper presented at the International Conference on Speech Prosody 2002, Aix-en-Provence, France.