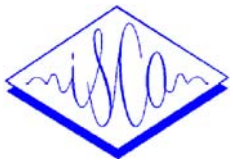

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Emotional Expression in Prosody: A Review and an Agenda for Future Research

Klaus R. Scherer & Tanja Bänziger

Department of Psychology
University of Geneva, Switzerland
klaus.scherer@pse.unige.ch

Abstract

This paper addresses the mechanisms underlying the effects of emotions on voice and speech, with a particular emphasis on intonation contours. After reviewing a number of conceptual issues, such as the different types of affective states, the nature of vocal affect communication, and the effects of push and pull factors on intonation, we describe an empirical study that examines statistically the existence of emotion-specific intonation contours by using a new coding system for the assessment of F0 contours in emotion portrayals. Throughout this paper, some suggestions for future work in this area are introduced.

1. Introduction

Ever since the teaching of rhetoric by Greek and Roman philosophers, the powerful effect of emotion on speech, with respect to both voice quality and prosody, has been highlighted. Although empirical research during the last decades has documented the acoustic correlates of different states of speaker affect, work on emotional expressivity in prosody has been somewhat neglected (but see [6]). There are many proposals for emotion-specific prosodic patterns in the literature; however most of them are based on selected examples rather than on systematic research. We suggest that progress in this research domain requires conceptual clarifications and the development of a research strategy that is designed to uncover the mechanisms underlying the apparent link between emotion and intonation.

We first need to define what, exactly, we mean by the term *emotion*. The unfortunate tendency is to use this term as a synonym for all kinds of speaker states that may have an affective element to them but that can hardly be considered to be full-fledged emotions. Scherer [20] has proposed a design feature approach to distinguish the following classes of affective states:

- Emotions (e.g., angry, sad, joyful, fearful, ashamed, proud, elated, desperate)
- Moods (e.g., cheerful, gloomy, irritable, listless, depressed, buoyant)
- Interpersonal stances (e.g., distant, cold, warm, supportive, contemptuous)
- Preferences/Attitudes (e.g., liking, loving, hating, valuing, desiring)
- Affect dispositions (e.g., nervous, anxious, reckless, morose, hostile)

The design features proposed for the differential definition of these states are partly based on a) response characteristics, such as intensity and duration or the degree of synchronization of different reaction modalities (e.g., physiological responses, motor expression, and action tendencies); b) antecedents (e.g., whether they are elicited by a particular event on the basis of cognitive appraisal); and c) consequences in terms of stability and impact on behavior choices. Table 1 shows a proposal for the specific feature profiles of each state (H - high, M - medium, L - low).

Table 1: *Types of affect*.

Types of Affect / Design features	Emotions	Moods	Interpersonal stances	Preferences Attitudes	Affect dispositions
Intensity	H	M	M	M	L
Duration	L	M	M	H	H
Synchronization	H	L	L	L	L
Event focus	H	L	M	L	L
Appraisal elicitation	H	L	L	L	L
Rapidity of change	H	M	H	L	L
Behavior impact	H	L	M	M	M

All of these states have been shown to affect voice and speech patterns, including intonation, even though there is currently little systematic empirical evidence. However, one can expect that the mechanisms that produce the effects are variable and may interact in complex ways for the different states. For example, each of these states is characterized by a specific pattern of interaction between "push effects" (the biologically determined externalization of naturally occurring internal processes of the organism, particularly information processing and behavioral preparation) and "pull effects" (socioculturally determined norms or moulds concerning the signal characteristics required by the socially shared codes for the communication of internal states and behavioral intentions) [17]. Given that the underlying biological processes are likely to be dependent on both the idiosyncratic nature of the individual and the specific nature of the situation, relatively strong interindividual differences in the expressive patterns will result from push effects. Conversely, for pull effects, a very high degree of symbolization and conventionalization, and thus comparatively few and small individual differences, are expected. With respect to cross-cultural comparison, one would expect the opposite: very few differences between cultures for push effects and large differences for pull effects. In consequence, systematic research on affective features of intonation needs to clearly distinguish between these types and explicitly focus on a particular category in order to identify the mechanisms involved. Little else other than the production of inconclusive or confusing results is to be gained by glossing over the

10.21437/SpeechProsody.2004-84

differences between emotions, moods, and interpersonal stances (or speaker attitudes).

In this paper we will focus on full-blown emotions, which, following the design features in Table 1, can be defined as episodes of massive, synchronous recruitment of mental and somatic resources to adapt to or cope with a stimulus event that is subjectively appraised as being highly pertinent to the needs, goals, and values of the individual. Given the powerful mobilization of the autonomous and somatic nervous systems, one can expect a powerful impact of push effects on voice and speech [18]. At the same time, emotions serve important adaptive functions in social interaction and communication, which leads to further enhancement of the emotion effects on speech and nonverbal behavior. This communicative function has led some theorists (e.g. Fridlund [7]) to the unfortunate claim that emotional vocalizations do not express emotions but just convey messages to others. Vocalizations do both of course, as do all sign processes, as shown in the classic Organon Model proposed by Bühler [4] (Fig. 1).

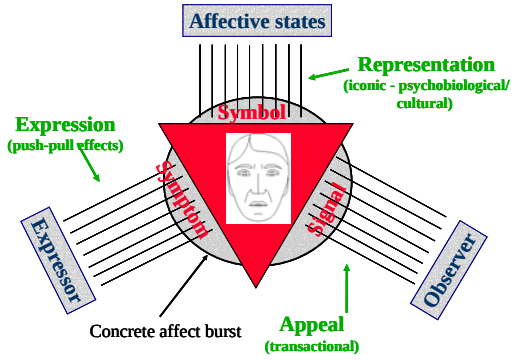

Figure 1: *Bühler's Organon Model.*

Bühler [4] proposed that each sign event has three simultaneous functions: it is a symptom of speaker state and thus expresses emotions, attitudes, and intentions; it serves as a signal to the perceiver or observer and constitutes an appeal to produce a reaction; and it is a symbol, serving a representational function that implies a shared meaning for the members of the respective culture. This triple function is important in that it underlines the multiple determinants that shape the expression pattern in a particular communication episode. In particular, the model highlights that expressive behavior is always shaped by both push and pull factors.

Scherer [21] has suggested that research needs to take this multiple determination into account by simultaneously focusing on both encoding and decoding of expressive signals and has proposed an adaptation of the Brunswikian lens model [3] for this purpose (see Fig. 2). However, the insistence on the joint operation of push and pull effects in any instance of expression does not mean that researchers should not try to disentangle these effects by appropriate research designs. Here, we strongly argue for research strategies that try to do exactly that because we are convinced that further progress in understanding the mechanisms underlying emotion effects on intonation could be greatly facilitated in this manner.

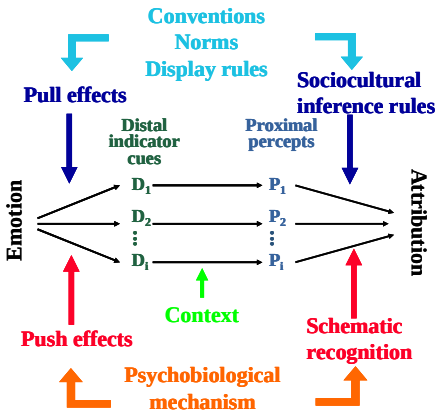

Figure 2: *Brunswikian lens model with push-pull effects.*

2. Studying push effects

On the basis of the physiological changes expected to occur with a number of major emotions, Scherer [18] has developed a set of predictions for the acoustic changes expected to accompany them. The main contributions to acoustic changes are modifications of respiration, muscle tension of the phonatory and articulatory apparatus, and vocal tract shape. The predictions concern mostly voice quality because it is difficult to make detailed predictions for intonation contours on the basis of physiological factors alone, except for general rising or falling due to changes in respiration. In order to test these predictions, we need to find human expressions that are primarily determined by push effects; that is, we need to find situations in which the sender does not engage in intentional communication and is relatively oblivious of any effects of an expression on others.

Assuming that there is evolutionary continuity in vocal affect expression [17], we can examine the evidence that there are push effects of emotional-motivational states on call contours in animal vocalizations. Morton [13] has suggested motivational-structural rules that seem to hold across many species of mammals. According to these rules, F0 of vocalizations increases with diminishing size and power of the animal, and the tonality of the sounds increases with increasing fear and tendency toward flight (see Fig. 3; line height indicates F0 mean, line direction indicates contour, line thickness indicates atonality, and vertical arrows signify that F0 could be lower or higher in the respective condition). As shown in Fig. 3, Morton assumes that vocalization contours are highly variable and that their shapes depend on particular contexts – in contrast to F0 mean and tonality, which can be predicted on the basis of the dimensions. If we were to translate these rules to emotional aspects of human utterances, we would expect F0 level to depend on submission/fear and spectral noise on aggression/anger. There would be no clear prediction for intonation contours except for rise-fall or fall contours in the case of moderate anger. What Morton's work clearly emphasizes is the need to distinguish different degrees of anger (horizontally) and fear (vertically) because the effect on vocalization may change dramatically. This is very much in line with Banse & Scherer's [1] insistence on the need to distinguish between hot and cold anger, anxiety and panic fear, or sadness and despair, each having very different vocal signatures. Because this important distinction is not often

made in the field, results tend to be equivocal and difficult to replicate.

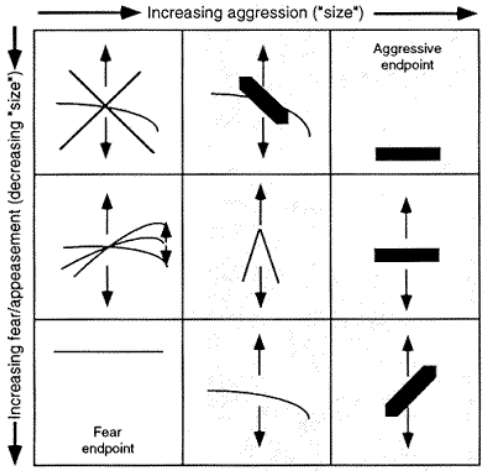

Figure 3: *Morton's motivational-structural rules.*

It should be noted that it is unlikely that animal calls are exclusively determined by push effects. Thus, Leyhausen [12] has shown that evolutionary pressure for "impression" (e.g., signal clarity) can affect expression patterns and Hauser [9] reviews extensive evidence showing that animals often manipulate their expressive behavior to produce a certain impression (pull effects). Yet, the study of animal vocalization provides one interesting avenue to examine the role of push effects on acoustic call structure, particularly because sophisticated experimental designs allow controlling the respective role of the two determinants [9].

A clear predominance of push effects can also be found in the human infant's grunts, sounds that are mere expressions of changing affective states, such as pain, hunger, and joy [9] (pp. 482-484). In contrast, infant babbling, laughing, and crying might already show rudimentary shaping by pull effects.

The closest equivalent to animal calls and infant grunts in adults is the type of interjection that Scherer [19] has called *affect bursts*. They constitute the extreme push pole of the continuum suggested earlier, being exclusively determined by the effects of physiological arousal. Their occurrence should be universal, but their form should be variable over individuals and situations. The term affect emblem is suggested to denote the extreme pull pole, brief facial/vocal expressions that are almost exclusively determined by sociocultural norms or models and that in consequence show a high degree of conventionality. One would expect a large number of intermediate cases between these two extremes, i.e., nonverbal facial/vocal expressions that are triggered by a particular affect-arousing event and that show at least some degree of direct physiological effect, but are at the same time subject to shaping by pull effects, as evident in control and regulation attempts.

Because affect bursts are the most direct phylogenetic equivalent to animal calls, the respective acoustic patterns should be directly comparable, including F0 contours. Unfortunately, little work has been done on this interesting type of vocalization, with the notable exception of Schröder [25], who reports an acoustic analysis of 28 German affect bursts (or rather, given that they were portrayed, emblems) for 10 emotions. The results confirm the general finding in the literature that F0 level rises with the degree of arousal implied by the affect state. Closer inspection of the data from the F0 contour coding [25] (Table 4) shows some remarkable correspondences with Morton's suggestions (Fig. 3). The contours for anger and threat seem to be similar to the low fear/high anger cells and the contours for worry similar to the low anger/high fear cells.

To study relatively pure push effects in emotional speech, one would need to record samples in which speakers are unlikely to control or regulate their vocal output in the interest of self-presentation [8] or for other strategic aims. This situation is rarely the case for the types of natural speech material used in the field. Mostly this material was obtained from media broadcasts, including journalists reporting emotion-eliciting events, affectively loaded therapy sessions, or talk and game shows on television. Apart from the strong probability that speech was highly monitored for its impression on potential listeners (pull effects), there are problems in determining the precise nature of the underlying emotion and the effect of regulation. Laboratory emotion induction has been rarely used in this area, with the exception of stress, given the ethical and practical difficulties of experimentally inducing strong and distinctive emotions. Most of the research has used actor portrayals, often applying Stanislaski techniques, in hopes that actors will reproduce push effects to achieve authenticity of their emotion portrayals.

In consequence, most of the empirical results in the area have been obtained on the basis of actor portrayals. A comprehensive summary [23] (Table 1) shows that there is relatively good agreement between the findings from different studies. Furthermore, many of Scherer's theoretical predictions are confirmed by the empirical results [1], [23]. Yet, one has to keep in mind that, given the predominant use of actor portrayals, we cannot be certain about the respective contributions of push and pull effects. In those cases in which Scherer's push predictions were not supported, there may have been a particularly strong influence of pull factors. Unfortunately, there has been little work on pull factors and thus it is difficult to identify the nature of this influence. In the next section we will examine this issue.

3. Studying pull effects

The first question to ask concerns the principles that underlie the operation of push factors. This strategy might enable us to venture some predictions that can then be empirically tested. We will start by identifying four major pull factors:

- One of the most ancient, and probably most important pull factors is the mimicking of push factors. This factor makes evolutionary sense because the purpose of deceptive communication is to fake the presence of a particular sender state that is normally expressed through push factors. Thus, as Hauser shows [9], because body size is correlated with pitch as a result of anatomic constraints (which represents a push factor), animals will develop morphological and behavioral means to lower pitch. It is most likely that actors being asked to credibly portray certain emotions will want to use exactly that strategy, i.e., produce a vocal portrayal that corresponds to what a person would show when in the grip of a strongly felt emotion. They have two ways to carry out

this strategy -- one is to rely on their personal observations of people affected by strong emotion and the other is to produce the respective emotion via Stanislaski techniques in oneself. In this case, our predictions would be the same as in the case of push factors, as described earlier, which would not yield very much with respect to intonation contours.

- The second major pull factor concerns customs or social conventions about how specific emotions ought to be vocally rendered. Most likely, the push effects that are regularly mimicked have also been conventionalized in this way. However, over time, conventions take on a life of their own and may start to strongly deviate from their origin. And in a long-term sociohistoric process, conventions can be created. One of the general misgivings about using actors for portrayals is indeed the possibility that they use cultural stereotypes, created through the tradition of drama schools or Hollywood films, which are no longer linked to push effects. It is difficult to develop detailed predictions as to the nature of these effects without being able to draw on extensive cross-cultural comparisons. Clearly, when conventions are similar across many cultures, they are most likely to mimic push effects. When they are different, one can expect pure pull effects developed by indigenous social conventions. Here we would expect typical intonation contours for particular cultures, as proposed by O'Connor & Arnold [14], Fonagy & Magdics [5] or Léon & Martin [11], for example. Unfortunately, we know of no systematic research in which actors from different countries have been asked to portray emotions and interpersonal stances to allow systematic comparisons of the intonation patterns and voice qualities.
- Another obvious candidate for pull effects is sound symbolism. A classic example is signal attack and decay. Steep attack of amplitude and pitch is typically perceived as aggressive, whereas soft functions are seen as weak and relaxed. If sound symbolism is universal, we would expect strong intercultural similarities. This similarity is indeed what is found for infant-directed speech in which there seem to be similar contour shapes for different languages for a number of messages such as approval, prohibition, attention, and comfort [9] (pp. 331-335). Hauser [9] (p. 485) provides another nice example for the acoustic structures of whistle tunes produced by shepherds in different countries to convey messages to their dogs. For example, there seem to be clear, universal contour structures for the commands of "fetch" and "stop". Again, much of sound symbolism could be based on push effects, as, for example, in Ohala's frequency rules [15].
- Other pull factors can be constituted by a coding system in associated channels or modalities. For example, smiling changes vocal tract shape and produces rather notable acoustic changes [26]. If politeness rules in a culture proscribe a lot of smiling, they will exert a strong pull on vocal production. The strongest pull factor for vocal paralinguistics is, of course, language itself. The intonation contours proscribed by language serve as pull factors for any kind of speech. The intonation contours that language may require are almost infinite because prosody is tied to the complexity of the underlying syntactical structure of an utterance. In addition, the existence of marking in an arbitrary signal system like language adds the possibility of using violations for communicative purposes. Thus, Scherer et al. [24] argued that the pragmalinguistic use of intonation follows configurational rules, as compared with the continuous coding one finds in affective signaling through F0 movements. For example, these authors showed that rising and falling contours take on different pragmatic meaning in Y/N and Wh questions.
- Finally, optimal sound transmissions in a particular habitat can serve as pull factors. Hauser [9] (p. 479) describes cases of species in which the body size-F0 link is apparently broken by the requirements to communicate in the forest or in the open savannah.

Much of the preceding discussion, was about conventionalization. It is, of course, the third function in Bühler's model described earlier. Although its importance for language and any arbitrary signaling system is usually acknowledged, often forgotten is the fact that conventionalization is equally important for nonverbal signaling of affect. The specificity of this function for this domain of communication might be that here push factors and pull factors largely overlap; i.e., phylogenetically continuous, universal, physiology-based expression systems have become conventionalized and represent an affect expression code. One of the major problems in the field is that this third function, and the large extent to which representation is shared among the members of a culture, has been only very rarely the central object of study. Much of the work on intonation contours, for example, is based on individual theorist's intuitions, on unsystematically selected tokens, often of anecdotal character, or self-produced examples. In order to better understand the nature of this shared code, we need a large and systematically constituted corpus of affect-related vocalizations and speech samples. Only such a massive corpus would allow us to statistically examine the similarities and differences of the use of certain contour shapes for certain types of affective states on the sender side and differential recognizability on the receiver side. Although it would be beneficial to have such a corpus that consisted of naturally occurring affective speech samples, the chances of realizing such a project seem rather low. If we could convincingly demonstrate that many pull factors actually mimic push factors, the use of actor portrayals would not pose major problems for the examination of the representation function.

In the following section, we present a study that examines the possibility of using an actor-generated corpus of emotion expressions to study the question of whether there are emotion-specific intonation contours.

4. Empirical assessment of F0 contours in emotion portrayals

4.1. Purpose

The purpose of the study was to examine whether we could find specific contour types for a number of basic emotions, using a large corpus that has been extensively studied for voice quality and aggregate F0 measures [1]. In addition, extensive judgment studies have demonstrated that the speech samples in the corpus are reliably recognized with an accuracy that matches the level generally found in the field. The first task was to develop a contour coding system that is amenable to quantitative statistical analysis.

4.2. Development of an appropriate intonation coding system

In our opinion, linguistic models of intonation are inappropriate 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 suited to describe variations of pitch involved in emotional communication. Results of past studies [10, 24] 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 expressed 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.

Consequently, we favored 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 were trying to communicate a variety of emotions. Identification of speech segments relying, implicitly or explicitly, on syntactic or semantic aspects was therefore impossible, and linguistic models of intonation were 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 & Ladd [16] 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.

4.3. Method

The corpus used in this study consisted of 144 emotional expressions, which were sampled from a larger set of emotional expressions described in detail by Banse & Scherer [1]. Expressions produced by 9 actors were selected. All actors pronounced 2 sequences of 7 syllables (1. "hät san dig prong nju ven tsi." 2. "fi gött laich jean kill gos terr") and expressed 8 emotions: cold anger ('irrit') and hot anger ('rage'), anxiety ('anx') and panic fear ('paniq'), sadness ('sad') and despair ('desp'), happiness ('joy') and elation ('elat'). F0 was extracted by autocorrelation using the speech analysis program PRAAT [2].

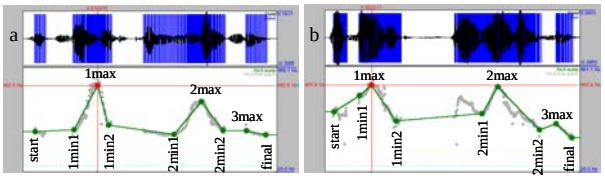

Figure 4: *Stylization examples.*

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, 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, minimum for the syllables "tsi" and "ter." Fig. 4 shows an illustration of this stylization for (a) a happy expression and (b) an expression of hot anger; both expressions are produced on utterance 1. The original F0 contours are represented by grey dots; the stylized contours are superimposed in green/black. 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. 4b. 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."

4.4. Results

The pattern represented in Fig. 4 – 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 was not affected by the expressed emotions, but varied for different speakers and for the two sequences of syllables that they pronounced (e.g., there were only 5 occurrences of the point '3min' for sequence 1 versus 42 occurrences of this point for sequence 2).

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. 5 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 expressed emotion.

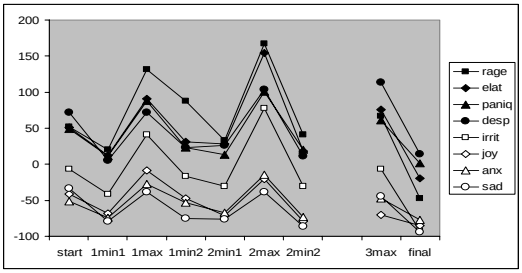

Figure 5: *Average F0 values by expressed emotion.*

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 shows that F0 level is mainly affected by emotional arousal. The F0 points for emotions with low arousal (such as sadness, happiness, and anxiety) are generally lower than the F0 points for emotions with high arousal (despair, elation, panic fear, and hot anger). 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) on the points represented in Fig. 5, there remains only a slight effect of expressed emotion on point '2max' and 'final.' The second maximum tends to be higher for recordings expressing elation, hot anger, and cold anger than for recordings expressing other emotions. The final F0 value tends to be relatively lower for hot anger and cold anger than for other emotions.

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 expressed emotion are presented in Fig. 6. The slopes tend to be steeper for part of the high-aroused emotions – especially for elation and hot anger – and less steep for part of the low-aroused emotions – especially for sadness, joy, and anxiety. 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 expressed 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. Differences remained essentially for the second F0 rise with, for instance, steeper slopes for elation, cold anger, and hot anger than for sadness and happiness, and for the final fall with, for instance, a steeper fall for hot anger than for the low-aroused emotions (happiness, anxiety, sadness, cold anger) and for panic fear.

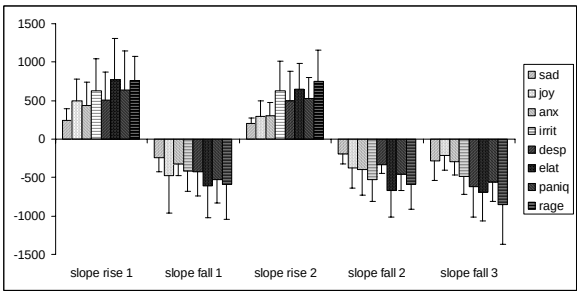

Figure 6: *Rising and falling F0 slopes, means, and standard deviations per expressed emotion.*

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 (calm joy) and the average location of the maximum F0 for elated expressions (aroused 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, panic fear, despair, and elation; although there was a significant decrease from the second local maximum to the third local maximum in all emotional expressions except for expressions of despair and elation. 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" stays higher in those expressions until the final fall than in expressions of cold anger and panic fear, which are also more "accentuated" on the second part than on the first part of the utterances. On the other hand, expressions of sadness, happiness, anxiety, and hot 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. 7 shows the means and standard deviations of this "declination" for each expressed emotion. The variance within expressed emotion being very important (see standard deviations in Fig. 7) and the number of expressions analyzed being relatively small, the differences between expressed emotions are not significant. Statistically, the F0 declination of expressions corresponding to hot anger (58 Hz per second) only tends to be steeper than the F0 declination of expressions corresponding to anxiety (16 Hz per second).

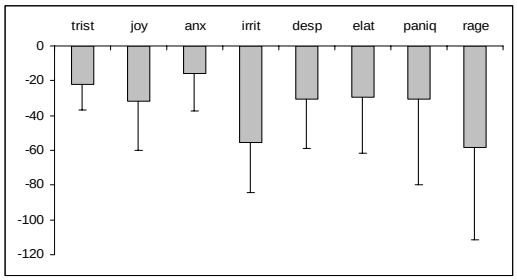

Figure 7: *Mean and standard deviation of F0 declination per expressed emotion.*

4.5. Discussion

The results presented earlier indicate that, in our corpus, expressed emotions affected mainly 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 expressed emotions.

However, a more detailed examination of the contours revealed specific differences for some expressed emotions. For some emotional expressions – especially hot anger, cold anger, and elation – the second F0 excursion in the utterances tended to be larger than for other emotions – such as sadness or happiness, 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.

Furthermore, the "shape" of the contours seems to be affected by the expressed emotions. Contours with "uptrend" shape (a term borrowed from Ladd et al. [10]) – 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 and elation, whereas expressions of sadness and happiness 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 expressed emotions. Emotions such as hot anger or elation might result in steeper final falls than expressions of anxiety or happiness.

The results regarding the relative height of local F0 excursions, contour "shape," and final fall must be considered with caution. The variations within expressed 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.

On the whole, then, there is little evidence for emotion-specific intonation contours. This is all the more remarkable, because we used nonlinguistic quasi-sentences that had no syntactic or semantic constraints or pull factors built in. 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 this short utterance may mean that push effects do not provide for contour coding other than the general level, range, and final fall parameters described earlier.

As mentioned earlier, this work certainly needs replicating 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. 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) [22] 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, emotion 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 mushrooming of emotion synthesis studies, few have advanced our knowledge. 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 and that we understand them because we experience them ourselves. Nothing could be further from the truth. And the vocal expression of emotion may be one of the most complex systems 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.

5. References

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

[2] Boersma, P.; Weenink, D. J. M., 1996. *Praat, a system for doing phonetics by computer, version 3.4* (132). Amsterdam: Institute of Phonetic Sciences of the University of Amsterdam.

[3] Brunswik, E., 1956. *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.

[4] Bühler, K., 1934. *Sprachtheorie* [Theory of language] (2nd ed.). Stuttgart, Germany: Gustav Fischer Verlag.

[5] Fonagy, I.; Magdics, K., 1963. Emotional patterns in intonation and music. *Zeitschrift für Phonetik* 16, 293-326.

[6] Frick, R. W, 1985. Communicating emotion: The role of prosodic features. *Psychological Bulletin* 97, 412-429.

[7] Fridlund, A. J., 1994. *Human facial expression: An evolutionary view*. San Diego: Academic Press.
[8] Goffman, E., 1959. *The presentation of self in everyday life*. Garden City, N.Y.: Doubleday Anchor.
[9] Hauser, M. D., 1997. *The Evolution of Communication*. Cambridge, Mass: The MIT Press.
[10] 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.
[11] Léon, P. R.; Martin, P., 1970. *Prolégomènes à l'étude des structures intonatives*. Montréal: Marcel Didier.
[12] Leyhausen, P. (1967). Biologie von Ausdruck und Eindruck. Teil 1. [Biology of expression and impression]. *Psychologische Forschung, 31*, 113-176.
[13] Morton, E. S., 1982. Grading, discreteness, redundancy, and motivation-structural rules. In *Acoustic communication in birds*, D. E. Kroodsma; E. H. Miller; H. Ouellet (eds.). New York: Academic Press, 182-212.
[14] O'Connor, J. D.; Arnold, G., 1973. *Intonation of colloquial English*. (2nd ed.). London: Longman.
[15] Ohala, J. J., 1994. The frequency code underlies the sound-symbolic use of voice pitch. In *Sound Symbolism*, J. Hinton; J. Nichols; J. J. Ohala (eds.). Cambridge, UK: Cambridge University Press, 325-347.
[16] Patterson, D.; Ladd, D. R., 1999. Pitch Range Modelling: Linguistic Dimensions of Variation. *Proceedings of the 13th International Congress of Phonetic Sciences*. San Francisco, 1169-1172.
[17] Scherer, K. R., 1985. Vocal affect signalling: A comparative approach. In *Advances in the study of behavior*, Vol. 15, J. Rosenblatt; C. Beer; M.-C. Busnel; P. J. B. Slater (eds.). New York: Academic Press, 189-244.
[18] Scherer, K. R., 1986. Vocal affect expression: A review and a model for future research. *Psychological Bulletin* 99, 143-165.
[19] Scherer, K.R., 1994. Affect bursts. In *Emotions: Essays on emotion theory*, S. van Goozen; N. E. van de Poll; J. A. Sergeant (eds.). Hillsdale, NJ: Erlbaum, 161-196.
[20] Scherer, K. R., 2000. Psychological models of emotion. In *The neuropsychology of emotion*; J. Borod (ed.). Oxford/New York: Oxford University Press, 137-162.
[21] Scherer, K. R., 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40, 227-256.
[22] 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.
[23] Scherer, K. R.; Johnstone, T.; Klasmeyer, G., 2003. Vocal expression of emotion. In *Handbook of the Affective Sciences*, R. J. Davidson; H. Goldsmith; K. R. Scherer (eds.). Oxford/New York: Oxford University Press, 433-456.
[24] Scherer, K. R.; Ladd, D.R.; Silverman, K., 1984. Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America* 76, 1346-1356.
[25] Schröder, M., 2003. Experimental study of affect bursts. *Speech Communication. Special Issue Speech and Emotion* 40(1-2), 99-116. Accessible at: http://www.dfki.de/˜schroed
[26] Tartter, V. C., 1980. Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics* 27(1), 24-27.