

Is there an emotion signature in intonational patterns? And can it be used in synthesis?

Tanja Bänziger^a, Michel Morel^b & Klaus R. Scherer^a

^a Department of Psychology, University of Geneva, Switzerland

^b CRISCO, University of Caen, France

Tanja.Banziger@pse.unige.ch

Abstract

Intonation is often considered to play an important role in the vocal communication of emotion. Early studies using pitch manipulation have supported this view. However, the properties of pitch contours involved in marking emotional state remain largely unidentified. In this contribution, a corpus of actor-generated utterances for 8 emotions was used to measure intonation (pitch contour) by identifying key features of the F0 contour. The data show that the profiles obtained vary reliably with respect to F0 level as a function of the degree of activation of the emotion concerned. However, there is little evidence for qualitatively different forms of profiles for different emotions. Results of recent collaborative studies on the use of the F0 patterns identified in this research with synthesized utterances are presented. The nature of the contribution of F0/pitch contours to emotional speech is discussed; it is argued that pitch contours have to be considered as configurations that acquire emotional meaning only through interaction with a linguistic and paralinguistic context.

1. Introduction

For several decades the role of intonation in the communication of emotion has been discussed. Some authors (e.g. Fonagy & Magdics [1]) have proposed that emotional meanings are conveyed by specific pitch contours and have provided examples to support their views. Using various types of speech manipulation techniques, other researchers have demonstrated that intonation contributes to the vocal communication of emotion. Classic studies in this direction have been carried out in the '60s by Lieberman & Michaels [2] and Uldall [3].

In 1984, Scherer, Ladd & Silverman [4] reported that two different mechanisms had been proposed in the literature to account for the way emotion affects speech: (a) Voice features can be seen as co-varying with the expressed emotional states (*covariation model*). Gradual modifications in different voice features (e.g. gradual increase in mean F0 or F0 range) would then, correspondingly, produce gradual modifications in the attribution of emotional states, independently of variations in other voice features. (b) Alternatively, different voice features and also linguistic features (semantic, syntactic or pragmatic) can be seen as combining in configurations (*configuration model*) to express emotions. In this view, the association between a set of voice features and a specific linguistic context would produce a specific emotional impression. Scherer, Ladd & Silverman [4] found support for both models in their study.

It is now rather well established that a number of acoustic features – including "overall" measures like F0 mean or range,

intensity mean or range, and speech rate – vary continuously with emotional arousal (see review by Scherer, Johnstone & Klasmeyer [5]). It is far less clear to what extent specific F0 contours can be associated with different emotions, especially independently of linguistic content. To examine this issue, quantifiable and comparable descriptions of F0 contours are needed. In this contribution results that have been obtained using a simple procedure to stylise F0 contours for emotional expressions are described. Emotions were portrayed by actors who pronounced sequences of syllables carrying no semantic or syntactic meaning.

In a second step, systematic variations have been applied to the F0 contours of 96 expressions generated with a TTS synthesiser. Aspects of the intonation (relative durations, F0 and intensity values) of 16 expressions produced by the actors were transferred to 16 new expressions generated with the same TTS synthesizer and were removed from the original recordings via resynthesis. Results from perception tests on the attribution of emotional meaning to the (re)synthesized recordings are presented.

2. Stylization of F0 contours

The corpus of emotional expressions and the procedure used for the stylization of the F0 contours are described below.

2.1. Method

The stylization has been applied to 144 emotional expressions that have been sampled from a larger set of emotional expressions described in detail by Banse & Scherer [6]. Expressions produced by 9 actors have been 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 has been extracted by autocorrelation using the speech analysis program PRAAT (Boersma & Weenink [7]).

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. 1 shows an illustration of this stylization for a happy expression (first utterance). The original F0 contour is represented by grey dots, the stylized contour is superimposed in green/black. Point eight ('3min') is missing in this expression.

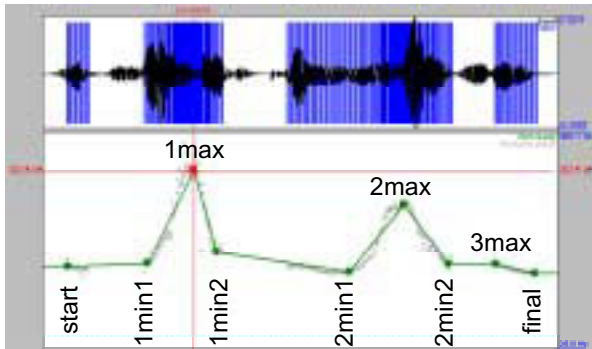


Figure 1: Stylization example

2.2. Results

The pattern represented in Fig. 1 – 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 sequence1 against 42 occurrences of this point for sequence2).

In order to control for differences in F0 level between speakers, a "baseline" value had to be defined for each speaker. An average F0 value was computed based on 112 emotional expressions (including the 16 expressions used in this study) produced by each speaker. Fig. 2 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.

Fig. 2 shows that F0 level is 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 graph 1, 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.

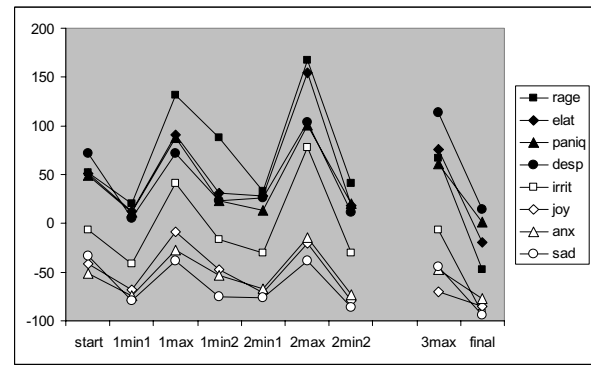


Figure 2: Average F0 values by expressed emotion

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 that there is a sizeable amount of variance around the average values shown for all measurement points.

3. Application to speech synthesis

Some aspects of the F0 contour – such as level, relative size of excursions, final fall – seem to be affected by expressed emotions. To test if such simple variations of the F0 contours can communicate emotional meaning in the absence of other vocal characteristics, different F0 contours have been applied on expressions produced with a TTS synthesizer (KALI, described in Morel & Lacheret-Dujour [8]). The possibility to perceive emotions in those expressions was tested in 2 judgment studies; the procedures and the results of those studies are described below.

3.1. Speech manipulations

The synthesized expressions that were used in this study have been created with the TTS system KALI according to the following criteria: The 2 sequences of syllables described above were used. A simple stylisation of the F0 contour (see Fig. 3) was used to produce two perceptible "accents" (a1 and a2) in each sequence. Two strengths were defined for the "accents", a 'weak' accent corresponding to an F0 excursion of 3 tones and a 'strong' accent corresponding to an F0 excursion of 6 tones. Two general levels were also defined, a 'low' level corresponding to the normal level for the synthetic voice, the 'high' level being raised with 4 tones. Three movements on the final syllable were defined as follows, a 'fall' (corresponding to a drop of 4 tones in the final syllable), a 'rise' (corresponding to a rise of 4 tones) and a 'flat' final end. In total 96 expressions were created: 2 synthetic voices (one male, one female) * 2 sequences of syllables * 2 levels * 2 accents * 2 accent strengths * 3 final ends.

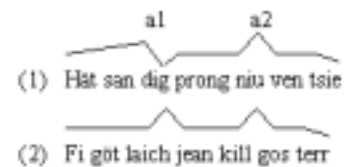


Figure 3: F0 stylisation applied to TTS synthesis

Aside from the systematic manipulation of F0 contours described above, an attempt was made to "cross-copy" the

intonation between a selection of 16 expressions (2 sequences of syllables * 8 expressed emotions) produced by the actors and a synthetic expression produced with the TTS system. On one hand, the F0 and intonation contours produced by the actors were transferred on the expression produced with the TTS system (creating 16 TTS generated expressions with "natural" F0 contours and synthetic voice) and, on the other hand, the F0 contour produced by the TTS system was applied to the 16 expressions produced by the actors (creating 16 resynthesized expressions with the voice quality of the actors and the intonation of the TTS system). All synthesized expressions are described more extensively in Morel & Bänziger [9].

3.2. Perception of emotionality and naturalness

Seven judges (1 male and 6 females, on average 32 years old) were recruited to assess the naturalness and the emotionality of the manipulated expressions and of the 16 original expressions (i.e a total of 144 expressions) presented in a different random order for each judge. Perceived naturalness (opposed to perceived artificiality) and emotionality (success versus failure to communicate an emotional impression) were judged on two continuous (visual analogue) scales ranging from 0 to 10. The scales were presented successively after each recording on a computer screen.

3.3. Results

Fig. 4 presents the average ratings for the original expressions (16 expressions produced by the actors), the systematically manipulated F0 contours applied to the female and the male TTS voices (96 recordings), the TTS expressions with the intonation of the expressions produced by the actors (16 recordings with emotional intonation but "neutral" voice quality) and the natural expressions resynthesized with the intonation of the TTS (16 recordings with emotional voice quality but "neutral" intonation).

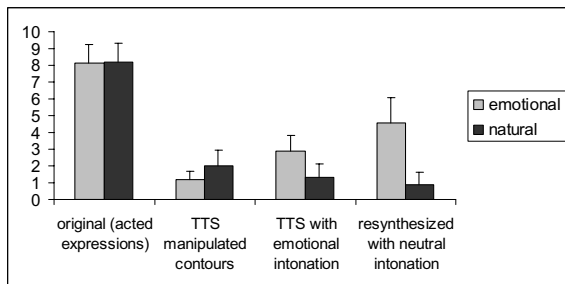


Figure 4: average ratings of naturalness and emotionality for different categories of expressions

The results mainly reflect a contrast between natural speech and manipulated speech. Nevertheless, a progressive increase in perceived emotional quality appears for the manipulated speech: the TTS generated expressions with the systematically manipulated contours are judged least emotional; the TTS generated expressions with the natural F0 contours receive higher emotional attributions; while the resynthesized expressions (with artificial F0 contour but emotional voice quality) receive the highest emotional attributions among the manipulated expressions. This difference in judgments of emotionality is not paralleled in the judgments of naturalness that show an opposed tendency: among the

manipulated expressions, the resynthesized expressions are judged least natural and the TTS expressions with manipulated contours are judged most natural.

With this judgment procedure, contrasting the manipulated expressions with natural expressions, no difference among the systematically manipulated F0 contours appeared. A second judgment procedure was used to test the possibility that different characteristics of the manipulated F0 contours could produce different emotional impressions.

3.4. Perception of specific emotional quality

A new group of 15 judges (13 females and 2 males, on average 25 years old) was recruited. Four continuous (visual analogue) scales representing the perceived intensity of 'anger', 'sadness', 'joy' and 'fear' were successively presented on a computer screen. The participants' task was to position icons representing the recordings on those scales. With this procedure, the participants can listen to the recordings as often as they wish by clicking on the icons displayed on the screen. Recordings can be directly compared and evaluated relatively to one another. The participants first rated the 48 TTS generated expressions with systematically manipulated F0 contours applied to the female voice. After a short break, the same participants rated the 16 resynthesized emotional expressions with the neutralized F0 contour and the 16 TTS expressions with the F0 contours copied from the actors' expressions. Ratings of the intensity of perceived 'anger', 'sadness', 'joy' and 'fear' had been obtained for the 16 original (actor produced) expressions using the same procedure in an earlier study.

3.5. Results

Fig. 5 shows that the original recordings received high ratings on the scales corresponding to the emotions expressed by the actors. The resynthesized recordings, on which a "neutral" F0 contour produced by the TTS system was applied, received slightly lower ratings on the respective scales. Especially the 2 recordings expressing 'elation' appear to communicate a much lower intensity of 'joy' after this speech manipulation. This drop in the perception of the expressed emotions is more important for the TTS recordings reproducing the F0 contours of the original expressions. It seems that only sadness is still recognized to some extent after this manipulation. These data suggest that F0 contour *in isolation*, independent of other affectively colored vocal cues, might often fail to communicate a specific emotional impression.

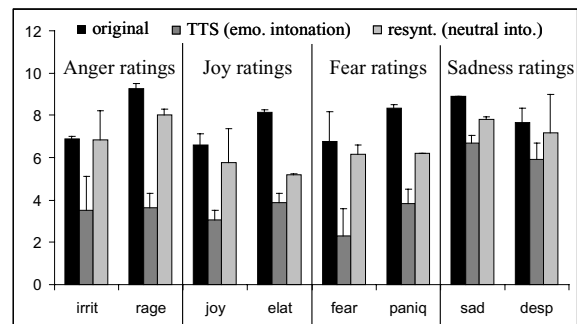


Figure 5: average intensity of perceived anger, joy, fear and sadness for 2 speech manipulations

For the systematically manipulated F0 contours applied to TTS recordings, the intensity judgments on the 4 emotional scales were mainly dependent on F0 level. A 'high' level of the global contour as compared with a 'low' level differentiated the perceived intensity mainly for 'fear' and 'joy', but also for 'anger' and 'sadness'. The average ratings for recordings with 'high' and 'low' level are presented in Fig. 6.

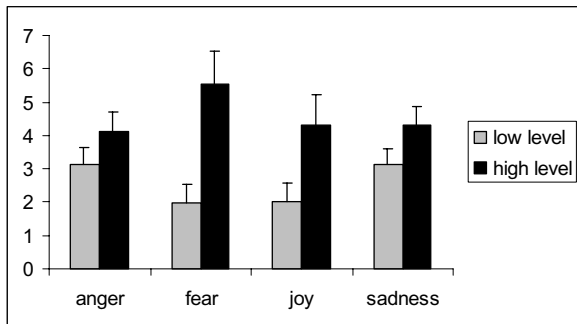


Figure 6: average emotional intensity ratings for TTS recordings with low and high F0 level

Based on the results described in section 2, we expected final pitch movement ('fall', 'rise', 'flat') and importance of the local F0 excursions ("accents") to affect emotional ratings. Data showed that recordings with a final rise received higher intensity ratings of fear and joy than recordings with a final fall. The size of F0 excursions also had small effects on certain aspects of the ratings. Thus, Fig. 7 shows that a small interaction effect between height of the second accent and overall F0 level appears for 'anger' ratings, whereas a small direct effect of the height of the second accent appears for 'joy' ratings.

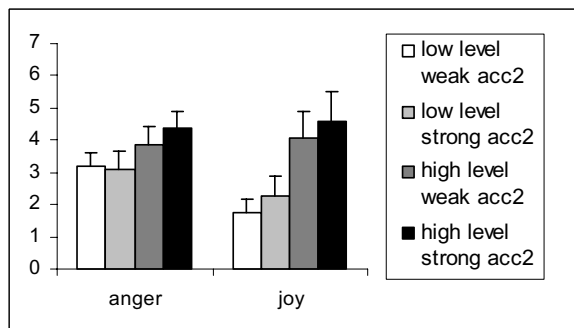


Figure 7: average joy and fear intensity ratings for expressions with high/low level and weak/strong acc2

4. Conclusion

The results from the analysis and the synthesis studies reported above show that only the overall level of the F0 contours was affected by expressed emotions and determined the emotion inferences of the judges in a powerful and statistically significant fashion. As could be expected from frequently replicated results in the literature [5], height of F0 seems to be reliably interpreted as indicative of differential activation or arousal. The current results do not encourage the notion that there are emotion-specific intonation contours. However, some of the detailed results suggest that aspects of

contour shape (such as height of selected accents and final F0 movement) may well differentially affect emotion inferences. However, it seems unlikely that such features will have a discrete, iconic meaning with respect to emotional content. It seems reasonable to assume that while the communicative value of F0 level may follow a covariation model, the interpretation of various features of F0 contour shape seems to be best described by a configuration model. Concretely, contour shape, or certain central features thereof, may acquire emotional meaning only in specific linguistic and pragmalinguistic contexts (including phonetic, syntactic, and semantic features), as well as normative expectations (e.g., in the sense of "given" vs. "new" [4]). Furthermore, the role of F0 contour may vary depending on the complexity of the respective emotion and its dependence on a sociocultural context. Thus, one would expect covariation effects for simple, universally shared emotions that are closely tied to biological needs, and configuration effects for complex emotions and affective attitudes that are determined by socioculturally variable values and symbolic meaning.

5. References

- [1] I. Fónagy and K. Magdics, "Emotional patterns in intonation and music", *Zeitschrift für Phonetik*, vol. 16, pp. 293-326, 1963.
- [2] P. Lieberman and S.B. Michaels, "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech", *Journal of the Acoustical Society of America*, vol. 34, pp. 922-927, 1962.
- [3] E. Uldall, "Dimensions of meaning in intonation". In D. Abercrombie, D.B. Fry, P.A.D. Maccarthy, N.C. Scott, and J.L.M. Trim, eds. *In honour of Daniel Jones: papers contributed on the Occasion of his Eightieth birthday*, pp. 271-279. London: Longman, pp. 271-279, 1964.
- [4] K.R. Scherer, D.R. Ladd, and K.E.A. Silverman, "Vocal cues to speaker affect: Testing two models", *Journal of the Acoustical Society of America*, vol. 76, pp. 1346-1356, 1984.
- [5] K.R. Scherer, T. Johnstone, and G. Klasmeyer, "Vocal expression of emotion". In R.J. Davidson, K.R. Scherer, and H. Hill Goldsmith, eds. *Handbook of affective sciences*, pp. 433-456. New York: Oxford University Press, 2003.
- [6] R. Banse and K.R. Scherer, "Acoustic profiles in vocal emotion expression", *Journal of Personality and Social Psychology*, vol. 70, pp. 614-636, 1996.
- [7] P. Boersma and D.J.M. Weenink, "Praat, a system for doing phonetics by computer, version 3.4", Institute of Phonetic Sciences of the University of Amsterdam, Report 132, 1996.
- [8] M. Morel and A. Lacheret-Dujour, "Le logiciel de synthèse vocale Kali : de la conception à la mise en œuvre". In Ch. D'Alessandro, ed. *Traitement Automatique des Langues n° 42*, pp. 193-221. Paris: Hermès, 2001.
- [9] M. Morel and T. Bänziger, "Le rôle de l'intonation dans la communication vocale des émotions : Test par la synthèse", *Cahiers de l'Institut de Linguistique de Louvain*, in press.