Analysis of Intonation Contours in Portrayed Emotions Using the Fujisaki Model

Maria O’Reilly, Ailbhe Ní Chasaide

Phonetics and Speech Science Laboratory
Centre for Language and Communication Studies
Trinity College Dublin, Ireland
{moreil12, anichsid}@tcd.ie

Abstract. This paper presents an analysis of f0 contours in portrayed emotions, using the Fujisaki model. The focus is on quantifying the f0 differences among the six emotions investigated (*surprised, bored, neutral, angry, happy,* and *sad*). The small dataset contained an utterance produced with the intention of portraying the six emotions (4 repetitions each). Preliminary results show that the Fujisaki parametrisation captures some striking intonational characteristics of these (intended) emotions. They indicate not only broad global differences, but also changes in the relationship of utterance internal constituents.

Keywords: Portrayed emotion, Fujisaki model, f0 contours, accent command amplitude and timing, declination.

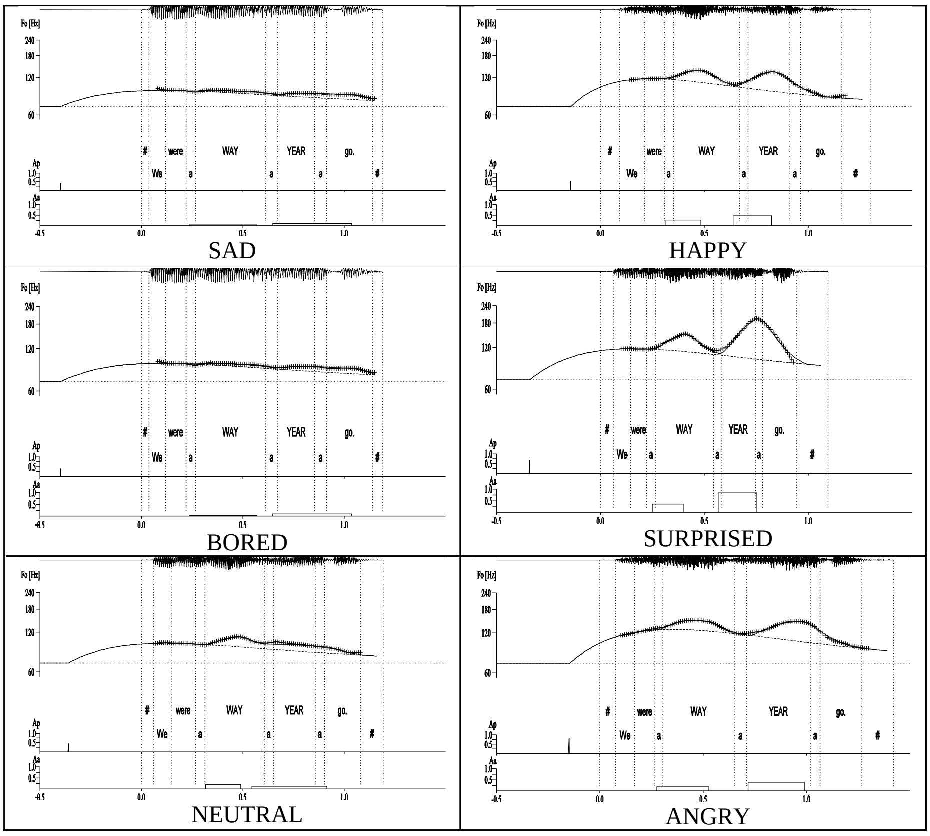
1 Introduction

It is widely understood that emotion is conveyed by means of prosodic parameters such as voice quality and the fundamental frequency [1], [2], [3], [4], [5]. In this paper the first aim was to capture, in quantitative terms, the salient intonational differences among repetitions of a semantically neutral utterance, produced so as to portray the following emotions: *surprised, bored, neutral, angry, happy,* and *sad.* To derive quantitative intonational measures we have used the Fujisaki model [6], [7]. The second aim of the study was to ascertain to what extent we can with this model glean an insightful account of the intonational differences among these portrayals. We are aware that other intonational correlates such as intensity, duration and voice quality may be crucial in further differentiating between emotions, however these are beyond the scope of this paper.

The present paper is furthermore intended to complement a currently ongoing analysis of the voice source correlates of the same dataset. Thus, although the present paper deals only with the f0 aspects, the ultimate intention is to piece together how the f0 and other dimensions of the voice source combine for the expression of affect.

The data of the present study are limited laboratory recordings of portrayed emotions, and thus can be criticised for not being necessarily indicative of what goes on in truly spontaneous, emotive utterances. However, we point out that within the obvious limitations that these data entail, the present approach may provide a useful complement to other approaches which are rooted in large and spontaneous databases.

Through resynthesis we hope further to explore the perceptual relevance of intonational and other source parameters in the perception of affect. And although it is quite likely that portrayed emotions are different in realisation from spontaneously occuring ones, there are many situations where portrayed emotions are in themselves of interest (e.g.: should we want to synthesise narrations of children's books) not as a substitute for true affect.



SAD

HAPPY

BORED

SURPRISED

NEUTRAL

ANGRY

Figure 1. Examples of the modelled f0 contours for six emotions. Upper panel displays (on a logarithmic scale in Hertz): speech wave, the original (crosses) and matched (solid line) f0 contours over Fb line with the time-aligned syllabic grid. Lower panel displays the phrase command (arrow) and accent commands (boxes) along a uniform time scale (-0.5 to 1.5 s).

1.1 The Fujisaki Model

The well-known Fujisaki model [6], [7] decomposes an f0 curve into a set of component curves which are attributed mainly to the activity of the cricothyroid muscle. The muscular activity involved in the f0 production is represented mathematically as a logarithmic sum of two time-varying terms plus a constant, f0 base value. All the Fujisaki parameters are briefly described below, and (apart from alpha and beta) presented in Figure 1.

Phrase Command (PC). This global component models the overall f0 trend at the IP/utterance level, and is related to declination. It assumes a gradual f0 drop over the course of an utterance. PCs are set in accordance with the syntactic structure of the sentence [6]. Declination of an utterance is modelled by at least one phrase command. The parameters that characterise a PC are its onset (T0), amplitude (Ap) and the natural angular frequency (alpha), which determines the rate of declination.

Accent Command (AC). This lower-level (local) term models the local f0 variations at the accent level (in relation to the accented syllables and prosodic boundaries). A linguistically-meaningful accent command is in our analysis assigned only where a prior auditory analysis determines that there is an accented syllable. The parameters pertaining to an AC are: the onset and offset of the accent command (T1 and T2), amplitude, or f0 peak scaling of the accent (Aa), and the natural angular frequency of the accent command mechanism (beta), which determines the rate at which the command reaches its maximum and then falls off. For dynamically-different accents in our dataset beta is allowed to vary.

Base Frequency (Fb). Fb is the f0 asymptotic value against which phrase commands are adjusted. It can be interpreted as an utterance-dependent parameter (allowed to vary between utterances) [6], [9], [10], or as a speaker-dependent parameter (kept constant) [8]. In the context of the present analysis, the approach chosen is the latter, as it was deemed desirable to constrain the degrees of freedom of the model. Consequently, any information concerning the average f0 dynamics (upper/lower register level) and the strength of the declination reset present in the intonation contours will be included in the phrase command component.

The adequacy of applying the Fujisaki model to pitch contours for emotional speech has been tested in recent studies. Higuchi and colleagues analysed f0 contours of Japanese in four speaking styles (*normal*, *kind*, *hurried* and *angry*) [9]. The parameters examined were the base frequency (varying), as well as phrase and accent command amplitudes, and all were found significant in distinguishing between the speaking styles examined (high Fb, very low Aa and Ap for *angry*, high Aa and lower Ap for *soft*, similar Ap and Aa in *hurried* compared to *normal* speaking style).

Hirose et al. presented a corpus-based method of f0 generation for Japanese in three emotions (*sadness*, *joy* and *anger*) alongside calm speech [10]. The findings relevant to this study include: no clear tendency for accent command and phrase command timings, the smallest Aa found for *sadness*, implying a flatter f0 contour and reduced dynamic range. Furthermore, the tendency in calm speech for reduced Aa in the course of an utterance was less evident in emotional speech. This was interpreted as an indication that the declination rate in emotional speech is slower.

2 Materials and Methods

The material chosen for this study is a set of 24 repetitions of the sentence "We were a**WAY** a **YEAR** ago", a typical declarative containing two accented syllables (capitalized in bold). There were 4 repetitions for each of the following emotions:

angry, *bored*, *happy*, *neutral*, *sad* and *surprised*. The informant was a male speaker of Hiberno-English, and the recordings were carried out in an anechoic recording room. The sentence was chosen as being rather neutral semantically, and because it contained mainly sonorants, which reduces the microprosodic influences.

The dataset was analysed with the use of PRAAT [11] and the FujisakiParaEditor [12], both freely available speech analysis tools. First, all repetitions were segmented in PRAAT in terms of syllables. The time-aligned syllable string allowed us to inspect the timing of the melodic contour relative to the syllabic tier as one of our intonational measures. F0 tracks were extracted with the PRAAT *To Pitch…* function, interpolated (to produce a continuous f0 contour for the subsequent parameter extraction), and finally smoothed. In cases where creaky voice was found, the f0 values were set to the values measured directly from the speech wave.

The second stage of f0 analysis involved the automatic extraction of the Fujisaki model parameters, and further fine-tuning by hand. To ensure consistency, the Fb value was set to a constant 70 Hz throughout. Initially, alpha was set to 2.0, while beta was set to 20.0. In order to provide a better fit to the data, manual fine-tuning was carried out, where both alpha and beta values were allowed to vary to capture more accurately the f0 contour shapes of the different sentences. The gamma value (ceiling level) for the accent command was kept constant at 0.9.

3 Results

Figure 1 illustrates how the f0 contours have been approximated with the Fujisaki model for each emotion portrayed. Phrase command and accent command parameters are presented in Tables 1 and 2, respectively. Figure 2 compares absolute and normalised Aa values. Figure 3 shows the timing of the accent commands relative to the accented syllable for the high-activation emotions only. While discussing the findings in the following sections, it has to borne in mind that because the dataset is relatively small, the observations should be regarded as tentative.

3.1 Phrase Command (PC) Parameters

As can be seen in Table 1, the main differentiating parameter for the phrase command is, as might be expected, its amplitude (Ap).

Ap. Phrase command amplitude distinguishes between affective contours at the utterance level. The high activation emotions (*happy*, *surprised* and *angry*) show increasingly high Ap levels. At the other end, the low activation emotions exhibit reduced Ap values, with *sad* reaching the lowest Ap.

T0. Phrase command onset was between 250 and 450 ms before the segmental onset of the utterance. No particular regularity in the behaviour of T0 was observed. For instance, the faster speech rate for *surprised* (see Figure 1) did not result in a shorter T0. The primary role of T0 can thus be seen as that of an anchoring point for the best possible matching phrase command.

Alpha. For most emotions, no strong relationship emerged for alpha (measure linked to the steepness of the declination line). With the exception of *happy* where a steeper declination slope is clearly present (nearly 2.6), alpha values for all other emotions were close to 2.0. We suspect that this parameter is not particularly meaningful in the low activation states - inspection of f0 contours in these emotions showed they exhibit little, if any, declination (especially in *sad*). As for the higher-activation states, alpha presents a certain degree of variation (*surprised* and *angry* with standard deviation of approximately 0.15). On the basis of the two facts we would conjecture that, like T0, alpha may not provide much information towards emotion differentiation.

Table 1. Mean values and their standard deviations for the phrase command parameters in six emotions. The means are given in bold type.

	SAD	*BORED*	*NEUTRAL*	*HAPPY*	*SURPRISED*	*ANGRY*
Ap	**0.43**/0.02	**0.47**/0.01	**0.47**/0.05	**0.52**/0.02	**0.66**/0.04	**0.78**/0.10
T0	**387**/87	**319**/57	**449**/12	**252**/44	**416**/33	**256**/34
Alpha	**1.90**/0.12	**2.00**/0.00	**2.03**/0.05	**2.58**/0.05	**2.10**/0.12	**2.00**/0.16

3.2 Accent Command (AC) Parameters

At the tone level, the crucial factor that characterises the emotions portrayed is the accent command amplitude. Additional information is given by the timing of the accent commands. Table 2 shows the results for AC parameters. Figure 2 illustrates the amplitude values in absolute terms, and normalised to the duration of the accented syllable. Figure 3 presents a way of considering how the melodic contour relates to the syllable tier, and shows the timing of the onset and offset of the accent commands (T1 and T2) relative to the beginning and end of the accented syllable. Accent command timing is only presented for the high activation emotions, given that in the lower affective states T1 and T2 are less likely to be meaningful (low Aa values).

Table 2. Mean values and their standard deviations for the accent command parameters in six emotions. The means are given in bold type. The numbers 1 and 2 relate to the first and second accent commands, respectively.

	SAD	*BORED*	*NEUTRAL*	*HAPPY*	*SURPRISED*	*ANGRY*
Aa1	**0.04**/0.02	**0.10**/0.02	**0.20**/0.02	**0.33**/0.05	**0.31**/0.04	**0.20**/0.04
Aa2	**0.09**/0.02	**0.07**/0.02	**0.12**/0.04	**0.46**/0.03	**0.82**/0.06	**0.40**/0.02
Dur1	**312**/40	**219**/10	**203**/20	**154**/10	**167**/30	**213**/30
Dur2	**393**/20	**334**/60	**325**/50	**210**/20	**203**/20	**288**/30
Beta1	**20**/0.0	**20**/0.0	**20**/0.0	**17**/2.0	**20**/4.3	**22**/1.5
Beta2	**19**/2.0	**20**/0.0	**20**/0.0	**16**/2.5	**16**/1.4	**14**/1.5

Aa. The very flat intonation contours of the low activation emotions are well captured by their low Aa values. As can be seen from Table 2, Aa values are lower for *sad* and *bored* than for *neutral*, and dramatically lower than for the three high activation emotions.

Figure 2 shows Aa both in absolute terms and normalised to the duration of the accented syllable. This was done as a way of ascertaining whether and to what extent higher Aa values might simply be correlated with (and “explained” by) the lengthening of the accented syllables for the stronger emotions. In the case of *angry*, the high Aa values are correlated with the longer duration, while for *surprised* the extremely high Aa values occur along with shorter syllable durations.

In the three strong emotions, *happy, surprised* and *angry*, it is clear that the two accented syllables are not equally boosted. The second (nuclear) accent becomes dominant, particularly in *surprised*. The consistency of this latter finding is indicated by the very low standard deviation for Aa values. This relative boosting of the nuclear relative to the prenuclear accent is striking for the strong emotions. It is worth noting that these changes in the internal relationships of the accented syllables are something that is lost when attention focuses only on global parameters.

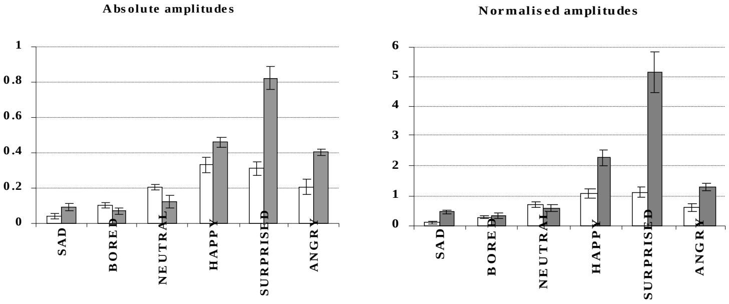


Figure 2. Mean values for absolute and normalised accent command amplitudes for first and second accent commands (white and grey bars). Whiskers represent standard deviations.

Duration. Accent command durations also appear to vary for the different emotions. *Sad, bored* and *neutral* exhibit longer durations, while *happy* and *surprised* have much shorter durations. We are not sure how much importance should be attached to the long accent commands of the low activation states. Given the very low amplitude of these commands, their timing may not be very critical to the modelling.

Beta. The value found for most accents is close to 20.0. In the lower activation states beta exhibits virtually no variation (except for the second beta in *sad*). This may, however, be an artifact of the modelling: accents with low Aa could be approximated equally well with a different beta value. Beta appears to be meaningful in the case of the high activation states (some variation confirms we are dealing with “true” accents), whereas it is not so for the lower activation states.

Relative timing of the accent command. Figure 3 shows the timing for the onset (T1) and offset (T2) of the accent command relative to the beginning and end of the accented syllable. As mentioned, this measure is presented for the strong emotions only, where Aa is substantial, and where T1 and T2 need to be precisely located to enable the Fujisaki modelling.

T1 is usually located (slightly) before the beginning of the accented syllable (for both the prenuclear and nuclear accents). T2 identifies the point at which f0 begins to drop, and for these utterances is more or less a measure of peak location. T2 exhibits more variation, and occurs somewhere in the middle of the syllable in the prenuclear accent, while it is timed later, closer to the end of the accented syllable in the nucleus. Nuclear postion is where T2 plays a differentiating role. While for *happy* and *angry* it is located just before the end of the accented syllable, for *surprised* it occurs at/just after the accented syllable boundary. The late peak in this rendition is quite audible.

Whereas the nuclear accent in the other cases would be heard as a falling (H+L) accent, in *surprised*, it is heard as a rise-fall (LH+L).

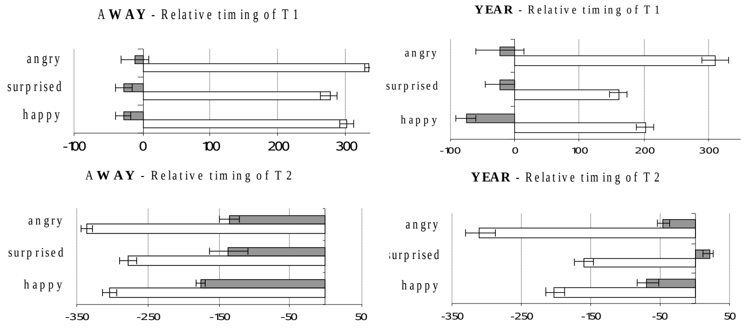


Figure 3. Mean values for accent command onset (T1) and offset (T2) (grey bars) relative to the accented syllable duration (white bars). The vertical line along 0 on the time axis represents the beginning of the accented syllable for T1, and the end of the accented syllable for T2. Standard deviations are shown as whiskers.

4 Discussion

These data do show that the measured parameters enable us to capture important intonational differences among the portrayed emotions. As for the phrase command, Ap does appear to differentiate between the high and low activation emotions. Additionally, alpha allowed further differentiation among the high activation emotions, in that the value for *happy* is much higher than for *surprised* and *angry*.

As expected, the accent commands differed considerably in terms of their amplitudes. Again, the high activation emotions have consistently higher Aa. It is striking for these high activation emotions that the second (nuclear) accent is the most dramatically altered, or, essentially upstepped, relative to the prenuclear accent. Thus the internal structure of the utterance is quite different from the neutral condition. Not only is the relative amplitude of the nuclear accent boosted (relative to the prenuclear), but the timing of the peak occurs later in the syllable. All of these effects are most exaggerated for *surprised*, where the very high nuclear peak is sufficiently delayed to be heard as a different nuclear contour.

5 Conclusions

These results are promising, showing that the quantification using Fujisaki parameters was indeed a fruitful method for characterising the intonational variations associated with the different emotions portrayed in these utterances. The parameters Aa and Ap were of particular importance in differentiation between at least the low and high activation emotions, as they effectively characterise variation in pitch range and dynamics. Alpha further served to differentiate among the high activation group.

Most research on f0 and emotion has focused on global measures such as f0 mean, range and dynamics. However, the present study highlights the fact that the internal relationships within the utterance are likely to be as crucial as the global measures. The upstep of the nuclear accent is a striking correlate of the high activation emotions studied here, particularly for *surprised.* Similarly, melodic alignment may be important to consider: here, the timing of the nuclear peak was found to vary, especially for *surprised*, where T2 differed from the other high activation emotions.

One of the advantages with the Fujisaki modelling is that the parameters are easily resynthesised. It is hoped that the perceptual relevance of these measurements can be ascertained through synthesis-based perception experiments.

As mentioned earlier, other prosodic dimensions also vary in the expression of emotion, real or portrayed. In a parallel study, these same utterances are being concurrently analysed in terms of the voice source parameters. Ultimately, we hope to pool this information to yield a fuller picture of the prosody of the voice.

Acknowledgments. The authors are grateful to Hansjörg Mixdorff for both the analysis software provided and for his generous assistance with its use. This work has been financially supported by the EU-funded Network of Excellence on Emotion, Humaine, and also by the research project, Prosody of Irish Dialects, funded by The Irish Research Council for the Humanities and Social Sciences.

References

1. Scherer, K.R.: Vocal Measurement of Emotion. In: Plutchik, R., Kellerman, H. (eds.): Emotion: Theory, Research and Experience, Vol. 4. Academic Press, San Diego, (1989) 233-259
2. Ní Chasaide, A., Gobl, C.: Voice Quality and f0 in Prosody: Towards a Holistic Account. Speech Prosody, Nara, Japan, (2004) 189-196
3. Campbell, N., Mokhtari, P.: Voice Quality: The 4th Prosodic Dimension. Proceedings of 15th ICPhS, Barcelona (2003) 2417-2420
4. Mozziconacci, S.: Speech Variability and Emotion: Production and Perception. Ph. D. thesis (1998) Technical University Eindhoven.
5. Bänziger, T., Scherer, K.R.: The Role of Intonation in Emotional Expressions. Speech Communication 46, (2005) 252-267
6. Fujisaki, H., Hirose, K.: Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese. Journal of the Acoustical Society of Japan (E) 5 (4), (1984) 233-241
7. Fujisaki, H., Ohno, S.: Prosodic Parametrization of Spoken Japanese Based on a Model of the Generation Process of f0 Contours. Proceedings of ICSLP-1996, Vol. 4. (1996) 2439-2442
8. Mixdorff, H.: Speech Technology, ToBI and Making Sense of Prosody. Speech Prosody 2002 Aix, France (2002) 31-38
9. Higuchi, N., Hirai, T., Sagisaka, Y.: Effect of Speaking Style on Parameters of Fundamental Frequency Contour. Proceedings of 2nd ESCA/IEEE Workshop on Speech Synthesis, Mohonk Mountain House, New Paltz, New York (1994) 135-138
10. Hirose, K., Sato, K., Asano, Y., Minematsu, N.: Synthesis of F0 contours Using Generation Process Model Parameters Predicted from Unlabeled Corpora: Application to Emotional Speech Synthesis. Speech Communication 46 (2005) 385-404
11. http://www.fon.hum.uva.nl/praat/
12. http://www.tfh-berlin.de/~mixdorff/thesis/fujisaki.html