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 above emotions (4 repetitions per emotion). 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 magnitude and timing, declination.

1 Introduction

It is widely understood that emotion is conveyed by means of a number of prosodic parameters such as voice quality, speech rate as well as the fundamental frequency [1], [2], [3], [4], [5], [6], [7]. 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 [8], and a second aim of this study was to ascertain to what extent we can with this model glean an insightful account of the intonational differences among these portrayals.

The present paper is furthermore intended to complement a currently ongoing analysis of the voice source correlates of these same utterances. Thus although the present paper deals only with f0/intonational 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 criticized 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 there are many situations where portrayed emotion is in itself desirable (e.g., narration of children's books) not as a substitute for true affect.

1.1 The Fujisaki Model

The well-known Fujisaki model [8], [9], [10] has been applied to many languages over a number of years. It decomposes a fundamental frequency curve into a set of component curves which are attributed mainly to the activity of two groups of the cricothyroid muscle which moves the thyroid cartilage relative to the cricoid cartilage, causing changes in the vocal cord length. The resulting sum in the vocal cord length is the sum of these two types of muscular changes (rotation and translation). The muscular activity involved in the f0 production is represented mathematically as a sum of two time-varying terms in the logarithmic scale plus a constant, F0 base value.

The Fujisaki model assumes a hierarchical order in prosodic structure of utterances: its components relate to the linguistic organization of an utterance as follows:

Phrase Command. This higher-level (global) component models the overall f0 trend at the intonation phrase/utterance level, and is related to declination. It assumes f0 gradually drops over the course of an utterance. Phrase commands are set in accordance with the syntactic structure of the sentence: they are reset when accompanying a respiratory pause, but they can also occur at major syntactic boundaries, and even at minor ones that are not followed by a pause [9]. Declination of an utterance is modelled by at least one (utterance-initial) phrase command. The phrase command contains three parameters pertaining to timing, amplitude and frequency information:

T0 and T0_e - the onset and offset of the phrase-level component

Ap - the strength (i.e. amplitude) of the declination reset

Alpha – the natural angular frequency of the phrase command mechanism, which essentially characterises the rate of declination.

Accent Command. This lower-level (local) term models the local f0 variations at the accent level. Accent commands serve to approximate the accentual patterns in relation to the accented syllables as well as (high) 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 accent command has the following parameters:

T1 and T2 – the onset and offset of the accent command. These timepoints are located at the start and the end of an f0 rise-fall pattern: T1 marks the beginning of the f0 rise towards the peak, whereas T2 is the beginning of the f0 fall. Depending on the structure of the phrase containing the accented syllable, one accent command can be associated just with the accented syllable, or stretch over a number of syllables..

Aa - the accent command amplitude. This parameter expresses numerical information about the f0 peak scaling of the accent.

Beta – the rate of an f0 rise-fall pattern, or – in other words – the dynamism of the f0 change in an accent. Beta can be interpreted as the factor determining the location of the f0 peak: for a higher beta the maximum accent command amplitude is reached more quickly (leftward peak shift), and also decays more quickly; the opposite is true for a lower beta value (rightward peak shift). We assume that – at least in the case of Hiberno-English – for dynamically-different accents beta is allowed to vary. Most importantly, beta and Aa changes are independent of the accent command timing: the accent command skew can be manipulated without affecting the command T1 and T2 anchor points.

Base Frequency, Fb. The third term of the Fujisaki model equation that can be interpreted in two ways. It is often treated as an utterance-dependent parameter allowed to vary (with the aim to produce the closest contour approximation) [9, 12, 13]. In an alternative view it can be interpreted as a speaker-dependent parameter which is kept constant [11]. In the context of the present analysis, the approach chosen is the latter, and differs from [12, 13]. This was felt to be desirable, both because it was thought likely to be truly a speaker-dependent constant, and furthermore 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 degree of final lowering present in the different portrayed emotions will be included in the phrase command component.

Summing up, an f0 contour is thus represented as the sum of three functions in a logarithmic scale: a slowly-decaying component (phrase command), local humps (accent commands) and a constant (Fb, or base frequency) over time. All the Fujisaki parameters, with the exception of alpha and beta, can be seen in Fig. 1 in the bottom panels of each modelled sentence.

The adequacy of applying the Fujisaki model to pitch contours for emotional speech has been tested in recent studies. Higuchi et al. analysed f0 contours of Japanese in four speaking styles (normal, kind, hurried and angry) [12]. The parameters examined were the base frequency, phrase command amplitude and accent command amplitude. These factors were found meaningful in conveying the differences between the speaking styles examined (high Fb and very low Aa and Ap for angry speech, high Aa and lower Ap for soft speech, similar Ap and similar (lower Aa) in hurried speech compared to normal speaking style). As mentioned, in the present analysis, the decision was made to maintain a constant base frequency, and therefore, results are not directly comparable.

Hirose and colleagues presented a corpus-based method of f0 generation for Japanese in three emotions (sadness, joy and anger) alongside calm speech [13]. Among the findings about the emotions that are relevant to this study are: no clear tendency for accent and phrase command timings, the smallest accent command amplitudes (Aa) were found for sadness, implying a flatter f0 contour and reduced dynamic range. Furthermore, the tendency in calm speech for the accent commands to reduce in amplitude 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.

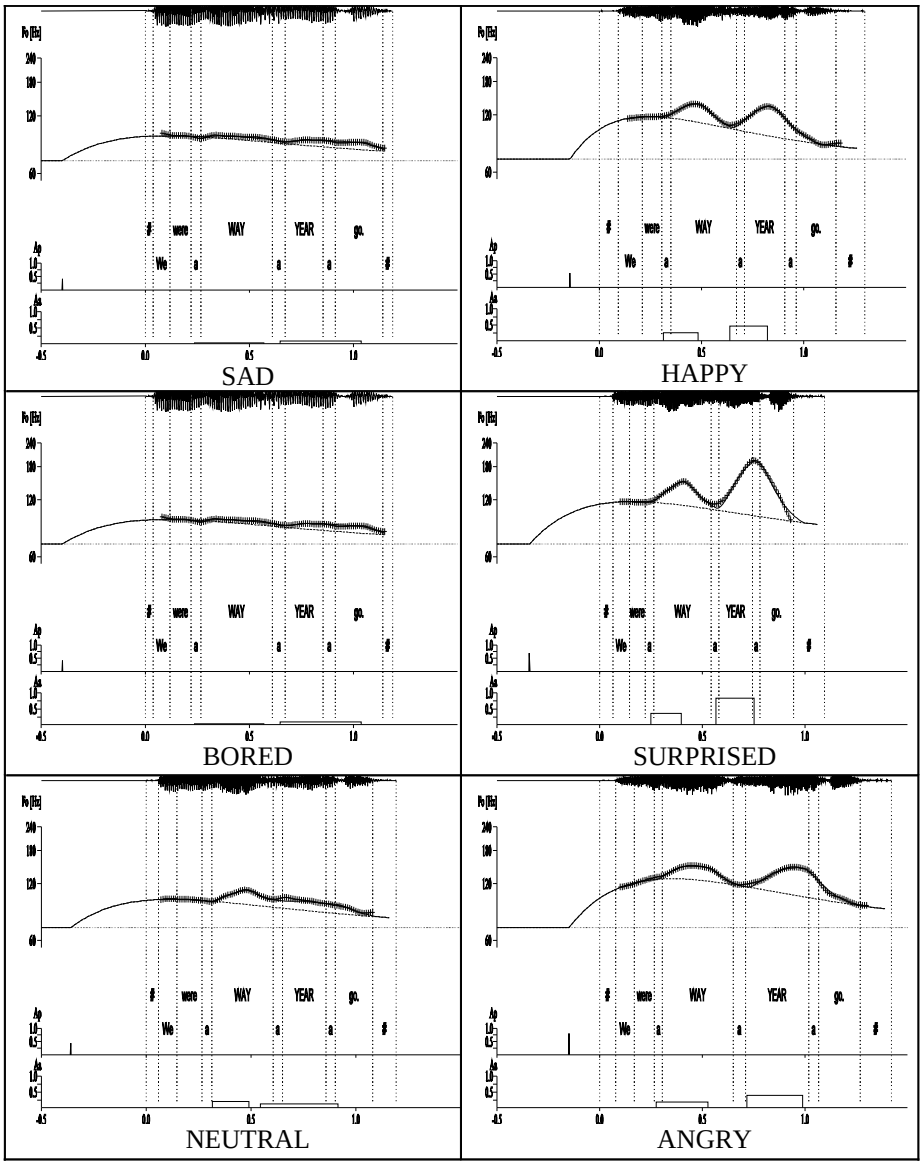


Fig.1. Examples of the modeled f0 contours for six emotions (from top to bottom): *sad*, *bored*, *neutral*, *happy*, *surprised* and *angry*. Upper panel displays: speech wave, the original (crosses) and matched (solid line – mostly identical to the extracted contour) f0 contours over the base frequency line with the time-aligned syllabic grid. The extracted and matched contours and Fb are shown on the F0 axis on a logarithmic Hertz scale. Lower panel consists of two parts: the phrase command onset and amplitude (arrow), and accent command onset, offset and amplitude (boxes) along a uniform time scale (-0.5 to 1.5 s).

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 as much sonorant material as possible something that minimise the microprosodic influences and the need for later smoothing procedures.

The dataset was treated in two stages with the use of two tools: PRAAT [14] and FujisakiParaEditor [15], both freely available speech analysis tools. First, all repetitions were orthographically labelled in PRAAT, and segmented 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 *To Pitch…* function for the 50-300Hz range, interpolated to produce a continuous f0 contour for the subsequent parameter extraction, and finally smoothed. In certain cases, namely where creaky voice was found, the fundamental frequency was measured directly from the speech wave, the pitch files edited, f0 points set, and only then interpolated and smoothed.

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 base frequency value was set to a constant 70 Hz throughout. Initially the alpha value of the phrase command was set to 2.0, while the beta value of the accent command was set to 20.0. Following the first, automatically derived estimates, manual fine-tuning was carried out to provide a better fit to the data, and here both alpha and beta values were allowed to vary to capture the characteristics of the different sentences. The gamma value (ceiling level) for the accent command was held constant at 0.9.

3 Results

Figure 1 shows for each emotion portrayed, a representative illustration showing how the f0 contour has been approximated by the Fujisaki model. Information concerning the phrase command measures are presented in Table 1, and the results concerning the accent commands are presented in Table 2 and Figure 2. Figure 3 shows for the accent commands the timing of the onset and offset (T1 and T2) relative to the beginning and end of the accented syllable. These timing measures are presented only for the high-activation emotions.

While discussing the findings in the following sections, it has to borne in mind that because the dataset is relatively small, the observations are tentative.

3.1 Phrase Command Parameters

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

Ap 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 emotion (*sad*) has a reduced Ap relative to *neutral.* For *bored,* Ap is the same as for *neutral.*

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*
T0_dist	**μ**	**387**	**319**	**449**	**252**	**416**	**256**
	σ	87	57	12	44	33	34
α	**μ**	**1.90**	**2.00**	**2.03**	**2.58**	**2.10**	**2.00**
	σ	0.12	0.00	0.05	0.05	0.12	0.16
Ap	**μ**	**0.43**	**0.47**	**0.47**	**0.52**	**0.66**	**0.78**
	σ	0.02	0.01	0.05	0.02	0.04	0.10

T0_dist. 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 distance. The primary role of T0 can thus be seen as an anchoring point for the best possible matching phrase command..

Alpha. For most emotions, no strong relationship emerged for alpha, which is a measure that is 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 in the low activation states is not particularly meaningful - inspection of f0 contours in these emotions showed they exhibit little, if any, declination (especially in *sad*). As for the higher-activation states, there is a certain degree of variation in the alpha values (*surprised* and *angry* with standard deviation of approximately 0.15). On the basis of the two facts we would conjecture that, like T0, the alpha measure may not provide much information towards emotion differentiation.

3.2 Accent Command Parameters

Table 2 shows the results for the accent command parameters. Figure 2 illustrates the amplitude values, first in absolute terms, and secondly, normalised to the duration of the accented syllable. Figure 3 presents a way of considering how the melodic contour relates to the syllable tier. It shows the timing of the onset and offset of the accent command (T1 and T2) relative to the beginning and end of the accented syllable, which was manually measured. These latter measures are only presented for the high activation emotions, given that in the other affects they are less likely to be meaningful, given the very low amplitude of the peaks.

Table 2. Mean values and their standard deviations for the accent command parameters (amplitude, duration and beta) 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*
Aa (1)	**μ**	**0.04**	**0.10**	**0.20**	**0.33**	**0.31**	**0.20**
	σ	0.02	0.02	0.02	0.05	0.04	0.04
Aa (2)	**μ**	**0.09**	**0.07**	**0.12**	**0.46**	**0.82**	**0.40**
	σ	0.02	0.02	0.04	0.03	0.06	0.02
Dur (1)	**μ**	**312**	**219**	**203**	**154**	**167**	**213**
	σ	40	10	20	10	30	30
Dur (2)	**μ**	**393**	**334**	**325**	**210**	**203**	**288**
	σ	20	60	50	20	20	30
β(1)	**μ**	**20**	**20**	**20**	**17**	**20.3**	**22.3**
	σ	0.0	0.0	0.0	2.0	4.3	1.5
β(2)	**μ**	**19**	**20**	**20**	**16.3**	**16**	**13.8**
	σ	2.0	0.0	0.0	2.5	1.4	1.5

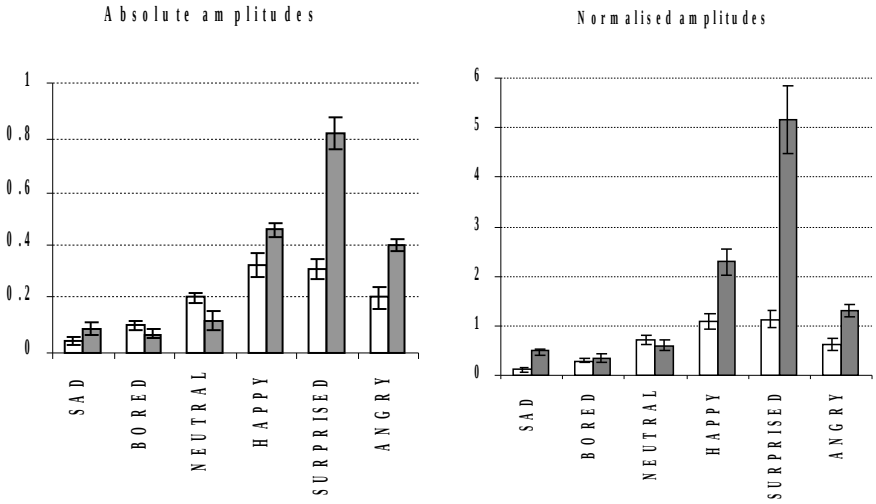


Fig. 2. Absolute and normalised accent command amplitudes (to the corresponding accented syllable durations). Mean values are presented as bars (white for the first and grey for the second accent command) and their standard deviations as whiskers.

Aa. The very flat intonation contours of the low activation emotions *sad* and *bored*, are well captured by their low Aa values. These are lower than the *neutral* values, and dramatically lower than those of the high activation emotions, *happy*, *surprised* and *angry*.

The normalised barchart in Figure 2, shows the Aa divided by 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 correlate with (and "explained" by) lengthening of accented syllables for the stronger emotions. The normalised figure shows that in the case of *angry*, the high Aa values are correlated with the longer duration, while for *surprised* the extremely high Aa values occur 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. The consistency of this latter finding is indicated by the very low standard deviation. This effect is most extreme in *surprised*. This relative boosting of the nuclear relative to the prenuclear accent is very striking for the strong emotions. It is worth noting that these changes in the internal relationships of the accented syllables is something that is lost when attention focuses only on global parameters.

Duration. The duration of the accent command does appear also to vary for the different emotions. *Sad*, *bored* and *neutral* have longer accent command 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 *sad*, *bored* and *neutral*. Given the very low amplitude of these commands, the timing of their onsets/offsets are not very critical to the modelling.

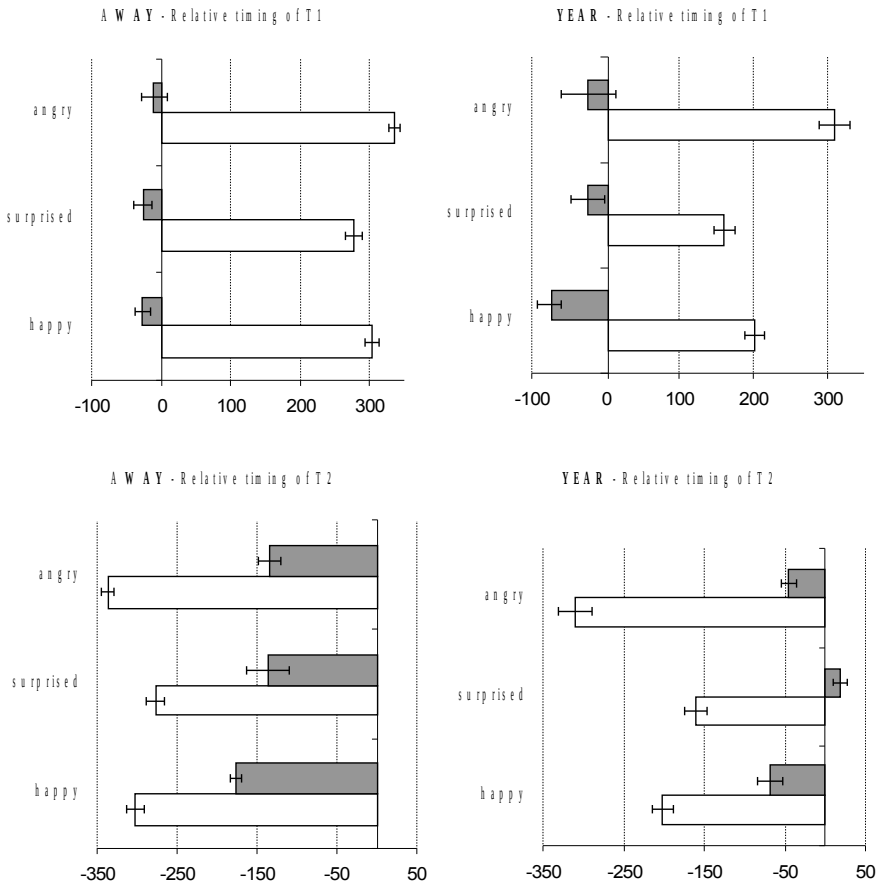
Beta. Beta values, which depict the steepness of the rise-fall tonal patterns, are close to 20.0 In the lower activation states they exhibit virtually no variation (except for the second beta in *sad*). This may, however, be an artifact of the modelling. Similarly to flat contours and their alpha, flat accents could be modelled with a different beta value, and the resultant match would be equally good. Beta is a meaningful factor in the description of the high activation states (some variation confirms we are dealing with "true" accents), whereas it is not so for the low ones.

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 only presented for the strong emotions, where the amplitude of the command is substantial, and where T1 and T2 need to be precisely located to enable the Fujisaki modelling.

With respect to the accent command onset (T1), it is usually located either at the beginning of, or slightly into the accented syllable (for both the prenuclear and nuclear accents). The offset, T2, identifies the point at which the f0 begins to drop.., and would be for these utterances more or less a measure of peak location. T2 is timed differently in the prenuclear (as compared to the nuclear) accent, occurring somewhere in the middle of the syllable. In the nucleus, T2 is timed later, closer to the end of the accented syllable. These differences are not relevant to the different emotions.

One timing feature however is important, i.e., the relatively late timing of nuclear T2 for *surprised*, where it occurs at or 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 (L+HL).

Figure 3. Duration of accented syllables; timing of T1 relative to the beginning of the accented syllable and timing of T2 relative to the end of the accented syllable. Mean values and standard deviations shown.



4 Discussion

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

As expected, the accent commands showed considerable differentiation in terms of their amplitudes. Again, the high activation emotions have strikingly higher amplitudes (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 is 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, and the alpha parameter 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. The amplitudes Aa and Ap are effectively characterising variation in pitch 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, and particularly for *surprised.* Similarly, local shifts in the timing of the melodic contour may be important to consider: in these utterances, the timing of the nuclear peak was found to vary, and is almost certainly a major correlate of these renditions of *surprised*, differentiating it from the other high activation emotions.

One advantage 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 in the introduction, other dimensions of the voice source 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 that we have used and for his generous assistance in showing us how Fujisaki modelling is carried out. 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. Ladd, D.R., Silverman, K., Tolkmitt, F., Bergmann, G., & Scherer, K.R.: Evidence for the Independent Function of Intonation Contour Type, Voice Quality, and F0 Range in Signalling Speaker Affect. Journal of the Acoustic Society of America 78 (2), (1985) 435–444
2. 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
3. Ní Chasaide, A., Gobl, C.: Voice Quality and f0 in Prosody: Towards a Holistic Account. Speech Prosody, Nara, Japan, (2004) 189-196
4. Paeschke, A., Sendlmeier, W.F.: Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements. ITRW on Speech and Emotion, Newcastle Northern Ireland, (2000) 75-80
5. Campbell, N., Mokhtari, P.: Voice Quality: The 4th Prosodic Dimension. Proceedings of 15th ICPhS, Barcelona (2003) 2417-2420
6. Mozziconacci, S.: Speech Variability and Emotion: Production and Perception. Ph. D. thesis (1998) Technical University Eindhoven.
7. Bänziger, T., Scherer, K.R.: The Role of Intonation in Emotional Expressions. Speech Communication 46, (2005) 252-267
8. Fujisaki, H.: 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 Press Ltd., New York (1988) 347-355
9. 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
10. 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
11. Mixdorff, H.: Speech Technology, ToBI and Making Sense of Prosody. Speech Prosody 2002 Aix, France (2002) 31-38
12. 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
13. 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
14. http://www.fon.hum.uva.nl/praat/
15. http://www.tfh-berlin.de/~mixdorff/thesis/fujisaki.html