QUANTITATIVE MODEL-BASED ANALYSIS OF $F_0$ CONTOURS OF EMOTIONAL SPEECH

Jesin James$^1$, Hansjorg Mixdorff$^2$, and Catherine I. Watson$^3$

$^1$,$^3$The University of Auckland, New Zealand

$^2$Beuth University, Berlin

jjam194@aucklanduni.ac.nz, mixdorff@beuth-hochschule.de, c.watson@auckland.ac.nz

ABSTRACT

Emotional speech recognition and synthesis are currently in focus due to emerging applications in Human-Computer Interaction. Past studies found the need to include secondary emotions (e.g. worried, apologetic) along with primary emotions (e.g. sad, happy), as social interactions exhibit subtle nuances. The $F_0$ contours of utterances spoken with five primary and five secondary emotions produced by two male and two female New Zealand English speakers were parameterized using the Fujisaki model. Automatic extraction followed by manual adjustments yielded amplitude and timing parameters on the utterance, phrase and syllable levels. Results show that the ten emotions significantly influence amplitude parameters on all levels, especially as a function of speaker-activation, being high in ‘excited’ as compared to ‘sad’, for instance. Also, the frequencies of accents and distributions of tonal transitions were analysed. As the Fujisaki model offers continuous $F_0$ contours, these results are directly applicable to emotional speech synthesis.

Keywords: Secondary emotions, Fujisaki Model, Fundamental frequency, Accent

1. INTRODUCTION

High-quality emotional speech syntheses are key components of state-of-the-art Human-Computer Interaction (HCI) applications. In social interactions, [1] has reported that emotions to be processed by such systems are not only the primary ones, such as happy, angry, sad, but also secondary emotions, e.g. worried, apologetic and anxious that are more subtle in nature. There have been studies that analyse the acoustic parameters of emotional speech to understand which features distinguish one emotion from another, and facilitate modelling these emotions for synthesis purposes (e.g.: [2–6]). In many studies Fundamental frequency ($F_0$) has been found to be a significant emotion-distinguishing factor (e.g.: [2–4]). The descriptive features of $F_0$ like mean and range provide useful information about emotion separation. But in speech synthesis the complete $F_0$ contour needs to be carefully modelled, as it conveys concurrent linguistic information - e.g. sentence modality and word prominences, as well as paralinguistic information (e.g. emotional variations). Hence, there is a need to parameterise the $F_0$ contour via a quantitative model. The Fujisaki model [9] parameterises the $F_0$ contour superimposing (1) the base frequency $F_b$, indicated by the horizontal line at the floor of the $F_0$ pattern of the utterance as shown in Figure 1, (2) the phrase component, the slowly drooping phrasal contours accompanying each prosodic phrase, and (3) the accent component, reflecting fast $F_0$ movements on accented syllables and boundary tones. The input functions of the model are impulse-wise so called phrase commands and rectangular accent commands which are displayed in the two lower tracks of Figure 1. These commands are specified by the following parameters: (1) Phrase command onset time $T_0$: Onset time of the phrasal contour, typically before the segmental onset of the ensuing prosodic phrase. (2) Phrase command amplitude $A_p$: Magnitude of the phrase command that precedes each new prosodic phrase, quantifying the amount of reset in the declination line. (3) Accent Command Amplitude $A_a$: Accent command amplitude associated with every pitch accent. (4) Accent command onset time $T_1$ and offset time $T_2$: The timing of the accent command that can be related to the timing of

Figure 1: Fujisaki model parameters for ‘Sound the horn if you need more’ (SAMPA phonetic symbol). $T_0$, $T_1$, $T_2$ are marked for first phrase and accent commands
the underlying segments. The model approximates the natural contour and interpolates through unvoiced sounds. The Fujisaki model is event-based, i.e. every command is related to the onset of a new phrase, accented syllable or boundary tone. It is parsimonious as large parts of the contour can be captured using few parameters and the parameters can be directly used to synthesise the contour. Due to these reasons this model has been used in this study. Past studies used the model in various prosodic contexts, e.g.: analysing tonal contrasts in Vietnamese [8], the influence of underlying attitudes and emotions in German [9] and Mandarin [10], as well as arousal in a psychotherapeutic setting [11] and perceived syllable prominence in German [12]. The use of the Fujisaki model to analyse emotional speech is a less explored area. [14] has analysed six emotions based on the Fujisaki model and [13] has employed the Fujisaki model for voice conversion for emotional speech. With growing needs in HCI, there is a need to also study some secondary emotions along with the primary emotions and quantitatively analyse them using the Fujisaki model. This is with the goal of synthesizing contours reflecting the emotion differences.

2. EMOTIONAL SPEECH CORPUS

To analyse the emotions needed for HCI, an open-source emotional speech corpus (available at: github.com/tli725/JL-Corpus) was developed, containing five primary (angry, happy, neutral, sad, excited) and five secondary emotions (anxious, apologetic, enthusiastic, worried) elicited by two male and two female professional New Zealand English speakers. It has strictly-guided simulated emotions with 15 sentences for each emotion. Fujisaki model fitting of the contour involves manual corrections, hence a subset of the corpus (50%) covering all emotions and all speakers equally was used. The corpus contains 2400 short utterances and 1200 were used. Hence, 120 sentences were analysed for each emotion. The sentences in the corpus are semantically neutral, except two emotionally coloured ones for each of the secondary emotions with semantics conveying that emotion (e.g. ‘I owe you an apology.’ for apologetic). The corpus was segmented at the syllable-level using HTK-based [16] American English forced alignment system and hand-corrected where needed.

3. FUJISAKI MODEL ANALYSIS

First, the contour was extracted using the Praat standard method [17] using a time step of 0.01s. Then the Fujisaki model parameters were estimated from the natural contour using an automatic algorithm [18]. In the analysis of reading-style speech typically every content word is characterized by at least one accent command associated with the primary pitch accent and the base frequency \( F_0 \) is kept constant for each speaker [19]. In the context of emotional speech, however, in principle every syllable can exhibit an accent command, especially when the emotion entails strong arousal. Sometimes even a single syllable that is strongly emphasized can contain two accent commands as seen in the syllable ‘m_o:’ of Figure 1. Among the total number of accent commands analysed 14.8% are cases where 2 accent commands are associated with a single syllable. We also observed that for certain speakers the \( F_0 \) has to be adjusted (±10Hz) as a function of the emotion portrayed. The Fujisaki model parameters for each utterance were checked to ensure that potential errors in \( F_0 \) tracking did not affect them, leading to additional accent commands in unvoiced segments. Finally, an automatic time alignment of the Fujisaki parameters with each of the syllables in the corpus was performed, i.e., accent commands are associated with the syllables in which they begin and end, and phrase commands with the initial syllables of phrases that they precede. The resulting database contains the Fujisaki model parameters (\( A_p, A_r, F_0, T_0, T_1, T_2 \)) for each of the syllables. In total the Fujisaki model-based parameterised data contains 4777 accent commands and 1701 phrase commands almost equally distributed among the ten emotions. As the \( F_0 \) contour can be described as a sequence of linguistically motivated tone switches, major rises and falls [19], the tone switches have also been marked based on the accent command alignment with respect to the syllable. Table 1 displays a list of accent command alignment options derived from [20] and their associated codes henceforth used in this paper. As a result, each syllable exhibits a switching code associated with it. In cases where there are two accent commands in one syllable (like ‘m_o:’ of Figure 1), combination tone switch codes were produced by concatenating the two codes associated with the two accent commands using a ‘+’ sign.

4. RESULTS AND DISCUSSION

Figure 2 shows parameterised \( F_0 \) contour results of the sentence ‘Find your boot is in this shoot’ produced with four emotions. The panels show from top

<table>
<thead>
<tr>
<th>Table 1: Tone switch codes used in this paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone Switch condition in current syllable</td>
</tr>
<tr>
<td>Rising tone switch</td>
</tr>
<tr>
<td>Falling tone switch</td>
</tr>
<tr>
<td>Rising, then falling tone switch</td>
</tr>
<tr>
<td>Falling, then rising tone switch</td>
</tr>
<tr>
<td>Accent command across syllable</td>
</tr>
</tbody>
</table>
to the bottom (1) speech waveform, (2) extracted $F_0$ contour (Hz) indicated by + signs and the Fujisaki model-based contour as a smooth continuous line, (3) underlying phrase command amplitudes ($A_p$) and (4) accent command amplitudes ($A_a$). The results of the Fujisaki parameters obtained for each emotion and their relation to emotion production are discussed here. It can be seen that high arousal emotions like angry and anxious exhibit larger accent commands compared to low arousal apologetic. The trends for $A_a$ and $A_p$ for all the emotions can be seen in the boxplots in Figure 3 for speakers female2 and male1. As the $F_0$ baseline for $F_0$ was kept constant within each speaker, $A_p$ and $A_a$ are good indicators of the emotion influence on the phrase and accent levels, respectively. Apologetic and sad exhibit the lowest $A_a$ and $A_p$, while excited has the highest values. $A_a$ is the best discriminator among the four features (seen in Table 2) which lists all emotion pairs and features that distinguish them based on ANOVA and pair-wise t-tests (both at significance level of 0.05). The utterance-wise mean duration of accent commands ($D$) was the least distinguishing feature. The $D$ of female2 does not vary as much as that of male1. This may be due to female2’s higher speech rate (0.31 syllables/sec). A weak, but significant correlation between $A_a$ and $D$ [Pearson’s $r = -.146$, $n = 4738$, $p < .01$] was found. This suggests that higher accent command amplitudes are associated with shorter accent command durations. The number of accent commands per second ($R$) was calculated by counting the number of accent commands in one sentence and dividing it by the total duration (in secs) of the sentence. $R$ also distinguishes between emotions, with apologetic having the lowest number per second, and excited having the highest. For lower arousal emotions like apologetic the speakers tend to produce a smaller number of accented syllables as observable in the examples in Figure 2. This will in turn affect $R$. The number of accented syllables does not match exactly the number of accent commands, as mentioned previously there were cases where an accented syllable was associated with two accent commands (14.8 %). None of the features were able to distinguish apologetic vs. sad and happy vs. anxious.

The secondary emotion closest to primary emotions excited and happy on the valence-arousal (VA) plane is enthusiastic (Valence level indicates the

<table>
<thead>
<tr>
<th></th>
<th>angry</th>
<th>anxious</th>
<th>apologetic</th>
<th>enthusiastic</th>
<th>excited</th>
<th>happy</th>
<th>neutral</th>
<th>pensive</th>
<th>sad</th>
</tr>
</thead>
<tbody>
<tr>
<td>anxious</td>
<td>$A_p, R$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>apologetic</td>
<td>$A_p, A_a$</td>
<td>$A_p, A_a, R$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>enthusiastic</td>
<td>$A_a$</td>
<td>$A_a$</td>
<td>$A_p, A_a, R$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>excited</td>
<td>$A_a$</td>
<td>$A_a$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>happy</td>
<td>$A_a$</td>
<td>$A_a$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a$</td>
<td>$A_p, A_a$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>neutral</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, D, R$</td>
<td>$A_p, A_a, D, R$</td>
<td>$A_p, A_a, D, R$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>pensive</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>sad</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>-</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>worried</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td>$A_p, A_a, R$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Features that distinguish the emotion pairs based on ANOVA and t-test results. $D$: Sentence-level mean duration of accent commands, $R$: Number of accent commands per second.

Figure 2: Fujisaki model parameterisation of $F_0$ contours for 'Find your boot in this shoot' in four emotions.
pleasantness of the voice ranging from unpleasant (e.g.: sad, fear) to pleasant (e.g.: happy, calm). Arousal level specifies the level of reaction to stimuli and range from inactive (e.g.: sleepy, sad) to active (e.g.: anger, surprise) [23]). The secondary emotion closest to angry is anxious and the secondary emotions closest to sad are neutral and apologetic. From Figure 3 it can be noted that the feature values for the emotions that are close to a primary emotion cover a similar area under the boxplot, with differences in the median line and overall range. So, feature values of the secondary emotions fall in between the extreme values of the primary emotions as expected due to their non-extreme positions on a VA plane [1].

Analysing the tone switches, the most frequent were falling tones (-1). All the initial syllable tone switches were falling. For the final syllable position, the falling tone (-1) and the rising then falling tone switch (-11) were the most common. Also, the higher arousal emotions are characterized by more occurrences of -11 (accent starts and ends in the same syllable) whereas lower activation ones exhibit more 11, i.e., more syllables are covered by one and the same accent command. This means that accents commands in high arousal emotions are shorter and accents more prominent, not only with respect to the F0 interval reflected by Ap, but also the concentration of the peak. In contrast, in lower arousal emotions more syllables are strung together between two accents, forming the hat-pattern [21]. An inspection of the tone switch codes (marked in red in Figure 2) reveals that despite the underlying linguistic information being maintained the same for all the emotions, the combinations of the tone switches may vary. Hence the sequence of code switches is useful information in F0 contour synthesis.

5. CONCLUSION

In this study five primary and five secondary emotions were analysed using the Fujisaki model. We were able to show that model parameters and two derived features are significantly influenced by the emotions. Using the superpositional approach we can now quantify this influence on the utterance, phrase and accent levels. Also, the accent command alignment with the syllable was analyzed and found to vary depending on the emotion type. Alignment codes are similar to tone labels in ToBI [22]. However, as the Fujisaki model captures the entire F0 contour and not just certain landmarks, it also facilitates direct synthesis of F0 contours which is one of the objectives of our study. Parameters can be predicted based on linguistic and paralinguistic context, such as positions of accented syllables and emotion type, to regenerate the contour during speech synthesis. We are aware that emotions cannot be discriminated exclusively by F0, but other acoustic features need to be taken into account as shown in [7]. Future work will therefore be dedicated to the development of a prediction model combining the F0 contours with other prosodic features.
REFERENCES


