Exploring Prosodic Features Modelling for Secondary Emotions Needed for Empathetic Speech Synthesis

Jesin James , Balamurali B.T * , Catherine Watson , Hansjörg Mixdorff

Posted Date: 3 January 2023

doi: 10.20944/preprints202301.0008.v1

Keywords: Secondary emotions; emotional speech synthesis; fundamental frequency contour; Fujisaki model; low-resource; empathetic speech

Article

Exploring Prosodic Features Modelling for Secondary Emotions Needed for Empathetic Speech Synthesis

Jesin James [1], Balamurali B.T [2,*], Catherine Watson [1] and Hanjörj Mixdorff [3]

[1] Department of Electrical, Computer, and Software Engineering, The University of Auckland; jesin.james@auckland.ac.nz; c.watson@auckland.ac.nz
[2] Singapore University of Technology and Design, Singapore
[3] Beuth Univeristy, Berlin; Hansjoerg.Mixdorff@bht-berlin.de
* Correspondence: balamurali_bt@sutd.edu.sg

Abstract: A low-resource emotional speech synthesis system for empathetic speech synthesis based on modelling prosody features is presented here. Secondary emotions, identified to be needed for empathetic speech, are modelled and synthesised in this paper. As secondary emotions are subtle in nature, they are difficult to model compared to primary emotions. They are also less explored, and this is one of the few studies that model secondary emotions in speech. Current speech synthesis research uses large databases and deep learning techniques to develop emotion models. There are many secondary emotions, and hence, developing large databases for each of the secondary emotions is expensive. This research presents a proof-of-concept using hand-crafted feature extraction and modelling of these features using a low resource-intensive machine learning approach, thus creating synthetic speech with secondary emotions. Here, a quantitative model-based transformation is used to shape the emotional speech fundamental frequency contour. Speech rate and mean intensity are modelled via rule-based approaches. Using these models, an emotional text-to-speech synthesis system to synthesise five secondary emotions - anxious, apologetic, confident, enthusiastic and worried is developed. A perception test to evaluate the synthesised emotional speech is also conducted.

Keywords: secondary emotions; emotional speech synthesis; fundamental frequency contour; Fujisaki model; low-resource; empathetic speech

1. Introduction

Text-to-speech synthesis is used extensively for human-computer interaction. In human-computer interaction, the synthetic speech produced by computer systems (such as conversation agents and robots) is modelled to be human-like. This humanness in the voice makes the technology more acceptable to users [1–3]. In this context, synthesising emotions as produced by humans in social situations is essential. Emotions are broadly classified into primary and secondary emotions[1]. There has been extensive research on primary emotions and methods to synthesise them (A detailed review of past research is provided in Section 2). The studies reported in [2,3] show that for the voice of a robot to be perceived as empathetic, not only primary emotions but secondary emotions are also essential. But one can expect that modelling secondary emotion is harder compared to modelling primary emotions. This is because secondary emotions are subtle compared to primary emotions. Also, lexical information needs to be supported by the appropriate prosodic component to enable people to

[1] Primary emotions are innate to support fast and reactive response. Eg: *angry*, *happy*, *sad*. Six basic emotions were defined by Ekman [4] based on cross-cultural studies, and the basic emotions were found to be expressed similarly across cultures. The terms 'primary' and 'basic' emotions are used in literature with no clear distinction defined between them. For this study, the definition of primary emotions as defined above is used to be in alignment with studies in emotional speech synthesis [5]. Secondary emotions are assumed to arise from higher cognitive processes based on evaluating preferences over outcomes and expectations. E.g., relief, hope [6]. This distinction between the two emotion classes is based on neurobiological research by Damasio [7].

 doi:10.20944/preprints202301.0008.v1

correctly perceive secondary emotions [8], i.e., the sentence for which emotional speech is synthesised has to be correctly modelled at the accent and phrase levels in alignment with what is being said.

Although there are many secondary emotions, the focus of this study will only be on secondary emotions that are needed for human-computer interaction, especially the ones that have been identified to be needed for an empathetic voice. This choice is based on studies on healthcare robots [2,3]. These studies analysed dialogs spoken by the healthcare robot during various scenarios like greeting the user, providing medicine reminders and guiding the user in tasks. This analysis was followed by a perception test, which suggested that human users perceived empathy in the voice that had secondary emotions. The secondary emotions identified based on the analysis and perception tests in the previously mentioned studies are: *anxious*, *apologetic*, *confident*, *enthusiastic*, *worried*. The same secondary emotions are modelled in this study. Studies on these specific secondary emotions are limited, and so are the number of databases available to analyse them. Therefore, rather than relying on large databases and deep-learning models built based on them, we have focused on understanding the impact of the secondary emotions on prosody features — specifically the fundamental frequency (f_0) contour. Hand-crafted-feature extraction was used to extract f_0 contour features. The features are then modelled to produce f_0 contours of secondary emotions. Two other prosody features, namely, speech rate and mean intensity, are modelled by rule-based methods. Modelling the secondary emotions by hand-crafted feature extraction on a relatively small database, as described in this paper, leads to developing a low-resource emotional speech synthesis system.

2. Past Studies on Emotional Speech Synthesis

A survey of studies focusing on emotional speech synthesis using various techniques from the 1990s is summarised in Tables 1 and 2. Only the most cited papers that provide a good understanding of the emotional speech synthesis techniques used during these years are reviewed here. The most cited paper before the 2000s [9] has more than 200 citations. Between 2000 to 2010, the most cited paper [10] also has more than 200 citations. And finally, the most cited paper [11] after 2010 also has more than 200 citations, indicating an increased interest in emotional speech synthesis in recent years. This section will help the readers understand the change in requirements for databases and resources over the years for emotional speech synthesis. In the review that follows, a comparison of the size of the speech database needed for each of the approaches is described. The comparison of the database size is made with the databases used in the latest studies (after 2010) as the reference. Studies after 2010 use databases with more than 100 hours of recordings [11,12], and this will be considered as a "large" database. The databases that only have recordings of all diphones of a language [13] will be considered a relatively small database, and others that contain more number of recordings will be considered medium-sized databases.

Studies in the 1990s used rule-based emotional speech synthesis [9,13] on a base voice[2] that was developed using formant/diphone synthesis. Diphone synthesis requires recordings of all diphones in a language. Formant synthesis models the human acoustic system without requiring a large database. Prosody features like fundamental frequency (f_0), duration and intensity were modelled by rule-based approaches. The emotion-based rules were derived by extracting these features from a small database for each emotion. The feature extraction used hand-crafted approaches, and the changes in features could be explained in terms of the change in emotions.

[2] 'base' speech synthesis system refers to the synthesis system that is built initially, which has no emotional modelling. The emotion-based modelling is built on this 'base' speech synthesis system.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

Table 1. Selected emotional speech synthesis techniques from the 1990s.

Speech synthesis method	Emotional speech synthesis method	Approach	Resources needed	Naturalness	Emotions modelled
1993 [13] Diphone synthesis	Rule-based	Emotion rules applied on speech synthesis systems	All possible diphones in a language have to be recorded for *neutral* TTS [1]. E.g. 2431 diphones in British English. An emotional speech database (312 sentences) to frame rules is needed	Average [2]	*neutral*, *joy*, *boredom*, *anger*, *sadness*, *fear*, *indignation*
1995 [9] Formant synthesis	Rule-based	Rules framed for prosody features such as pitch, duration, voice quality features	DECtalk synthesiser used containing approximately 160000 lines of C code. Emotion rules framed from past research	Average	*anger*, *happiness*, *sadness*, *fear*, *disgust* and *grief*
2004 [14] Parametric speech synthesis	Style control vector	Style control vector associated with the target style transforms the mean vectors of the *neutral* HMM models	504 phonetically balanced sentences for average voice, and at least 10 sentences of each of the styles	Good	Three styles: *Rough*, *Joyful*, *sad*
2006 [10] Recorded *neutral* speech used as it is	Rule-based	GMM [3] and CART [4] based models for f_0 and duration	Corpus with 1500 sentences	Average	*neutral*, *happiness*, *sadness*, *fear*, *anger*
2006 [15] Parametric speech synthesis	Corpus-based	Decision trees determine f_0 contours & timing trained from the database	11 hours (excluding silence) of *neutral* sentences + 1 hour emotional speech	Good [5]	Conveying bad news, yes-no questions
2006 [15] Parametric speech synthesis	Prosodic phonology approach	ToBI [6] based f_0 modelling	11 hours (excluding silence) of *neutral* sentences + 1 hour emotional speech	Good	Conveying bad news, yes-no questions

[1] Text-To-Speech. [2] compared to the methods that were developed in the following years. [3] Gaussian Mixed Model, [4] Classification And Regression Tree. [5] reduced due to over-smoothing of spectral and excitation parameters by HMM models. [6] Tones and Break Index.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

Table 2. Selected emotional speech synthesis techniques from 1990s (Continued).

Speech synthesis method	Emotional speech synthesis method	Approach	Resources needed	Naturalness	Emotions modelled
2007 [16] Parametric speech synthesis	Model adaptation on average voice	Acoustic features Mel-cepstrum & log f_0 were adapted	503 phonetically balanced sentences for average voice, and at least 10 sentences of a particular style	Good	Speaking styles of speakers in the database
2010 [17] *Neutral* voice not created	HMM-based parametric speech synthesis	Each emotion's database was used to train emotional voice.	Spanish expressive voices corpus - 100 mins per emotion	Good	*happiness*, *sadness*, *anger*, *surprise*, *fear*, *disgust*
2017 [12] Parametric speech synthesis using recurrent neural networks with long short-term memory units	Emotion-dependent modelling and unified modelling with emotion codes	Emotion code vector is input to all model layers to indicate the emotion characteristics	5.5 hours emotional speech data + speaker independent model from 100 hours speech data	Reported to be better than HMM-based synthesis	*neutral*, *happiness*, *anger*, and *sadness*
2018 [11] Tacotron-based end-to-end synthesis using DNN [3]	Prosody transfer	Tacotron model learning a latent embedding space of prosody derived from a reference acoustic representation containing the desired prosody	English dataset of audiobook recordings - 147 hours	Reported to be better than HMM-based synthesis	Speaking styles of speakers in the database
2019 [18] Deep Convolutional TTS	Emotion adaptation	Transfer learning from *neutral* TTS to emotional TTS	Large dataset (24 hours) *neutral* speech + 7000 emotional speech sentences (5 emotions)	Reported to be better than HMM-based synthesis	*anger*, *happiness*, *sadness*, *neutral*

In the early 2000s, the trend shifted to parametric speech synthesis, with HMM (Hidden Markov model) based synthesis being the most popular (See rows 3 to 8 of Tables 1 and 2). Parametric speech synthesis increased the need for good quality databases[4] with adequate phonetic coverage (between 500 [15,16] to 1500 [10] sentences and larger corpora with 11 hours of *neutral*[5] speech recording [15]). Emotions were imparted to the synthesised speech using rules [10], where the rules were derived from a small corpus of each emotion. Corpus-based modelling [15] was also done, where an emotional speech corpus (one-hour recording for each emotion) was used to derive models for each emotion's prosody features. Another approach modelled emotional prosody phonology using ToBI (Tones and Break Index)-based f_0 contour modelling [15]. One-hour recording for each emotion and 11 hours of *neutral* speech recording were used. Another approach was style adaptation using HMM-based synthesis. Style control vectors [14] and adaptation of acoustic features like Mel-cepstrum and log f_0 were used. These adaptation methods need a relatively large database (approximately 500 phonetically balanced sentences [14,16]) to produce an average voice and a smaller database - approximately 10 sentences for

[4] The term good quality here refers to recordings in recording studio environments that have controlled noise levels.
[5] *Neutral* in this context refers to speech without any emotions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

each emotion/style to be adapted [14,16]). All these approaches used the HMM-based synthesis to produce a *neutral* voice and some form of emotion modelling to incorporate emotions onto the *neutral* voice. This required a medium size[6] database of *neutral* speech, and a small database of emotional speech to learn from. The features modelled using these approaches are interpretable. However, the naturalness of these synthesised voices was reported to be inferior due to the inherent disadvantage of HMM-based synthesis, that it over-smoothens the spectral and excitation parameters [19]. If a large database for each emotion (100 minutes of recordings per emotion) is available, an HMM-based synthesis can be achieved by training the models based on each of the emotional databases [17] without the need for a *neutral* voice. Such modelling of individual emotions will often produce emotional speech with good naturalness. However, developing large databases for each emotion will include too much overhead, such as the additional requirement to produce recordings for each emotion separately.

Post-2015, the trend in speech synthesis shifted towards incorporating neural networks in parametric speech synthesis (See Tables 1 and 2 Rows 9-11). Earlier approaches focused on using recurrent neural networks with long short-term memory units [12]. Then, emotion-dependent modelling was done by inputting emotion code vectors to all model layers based on an emotional speech database. The *neutral* voice was trained using a large database of 100 hours of speech, and the emotional speech data was 5.5 hours long. The speech produced by such deep neural networks is more natural than the HMM-based approach, as the over-smoothing of spectral and excitation parameters is avoided. With the improvements in neural networks, the availability of large databases and increased processing power, there has been a lot of focus on developing end-to-end emotional text-to-speech synthesis systems. Tacotron-based end-to-end speech synthesis system is one of the latest speech synthesis techniques. The study reported in [11] used Tacotron and implemented prosody transfer for emotional speech synthesis. Another research [18], used a deep convolutional neural network TTS and performed emotion adaptation via transfer learning. Both these neural network-based approaches require large databases (147 hours [11], 24 hours *neutral* speech + 7000 emotional speech sentences [18]). The interpretation of the features learned by the neural network is not directly possible. Rather, the learning is based on spectrograms and image-related features, and these features cannot be easily associated with the acoustic correlates of speech production.

The last column of Tables 1 and 2 lists the emotions that have been synthesised by these aforementioned approaches. Most of the emotions synthesised are primary emotions such as angry, sad, happy, fear (in studies [10,17,18,20]) and others focusing on speaking styles (like studies in [11,21,22]). Only a few studies [13] have synthesised some secondary emotions.

2.1. Emotional Speech Corpus

The JLCorpus[7] [23] contains a total of 2400 emotional speech sentences from 10 emotions (5 primary emotions - *angry*, *excited*, *happy*, *neutral*, *sad*; and 5 secondary emotions - *anxious*, *apologetic*, *confident*, *enthusiastic*, *worried*) spoken by two male (male1 and male2) and two female (female1 and female2) speakers of New Zealand English. The emotions in the JLCorpus on a valence-arousal [8] plane is shown in Figure 1. The JLCorpus is used for this study.

---

6 medium in comparison to the databases needed for the deep-learning approaches explained in the next paragraph

7 https://www.kaggle.com/tli725/jl-corpus

8 Valence indicates the pleasantness of the voice ranging from unpleasant (e.g. *sad*, *fear*) to pleasant (e.g. *happy*, *hopeful*). Arousal specifies the reaction level to stimuli ranging from inactive (e.g. sleepy, *sad*) to active (e.g. *anger*, *surprise*). Russel developed this model in a psychology study where Canadian participants categorised English stimulus words portraying moods, feelings, affect or emotions. Later, 80 more emotion words were superimposed on Russel's model based on studies in German [24]. Russel's circumplex model diagram, as used in this study shown in Figure 1, is adapted from [25], which was adopted from Russel and Scherer's work, but the positive valence is depicted by the right side of the x-axis (in contrast to Scherer's study where it was on the left side.). A two-dimensional model is used (and not higher dimension models) as representing the emotions on a plane facilitates their visualisation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

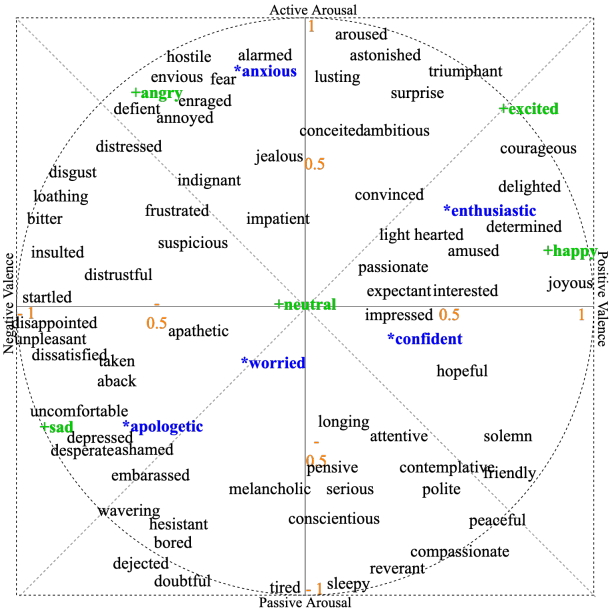

Figure 1. Emotions in the JLCorpus (blue * are the secondary emotions, and green + are the primary emotions) positions on the valence-arousal plane.

With the motivation to synthesise secondary emotions using a low-resource approach and hand-crafted features, the research questions in this investigation are:
Research Question 1: Can prosody features be used to model secondary emotions?
Research Question 2: How can a low-resource emotional text-to-speech synthesis system be developed for secondary emotions?

3. Emotional Speech Corpus Analysis

Even though the JLCorpus contains both primary and secondary emotions, the focus of this research is only on secondary emotions. The primary emotions *sad*, *excited* are added in the plots to represent the extremities in the valence-arousal levels, and *neutral* emotion is added as a baseline for comparison. This will help the reader understand the relative behaviour of the secondary emotions compared to the primary emotions. The analysis presented focuses on three prosody features - fundamental frequency (f_0), speech rate, and mean intensity. These three features were considered as they have been extensively used in past research (examples can be found in Tables 1 and 2) for emotional speech synthesis. Only the results of male2 and female2 speakers are discussed as they had the highest perception accuracy among all four speakers from the perception test for evaluating the JLCorpus [23]. Averaging the results across all the speakers will cause these feature values to not have distinct emotion-dependent regions. Therefore, speaker-based results averaged across the sentences per speaker per emotion are presented here. There are 60 sentences per speaker for every emotion; in total, the results correspond to 960 sentences. When the valence-arousal levels are described to relate to change in prosody features, the two-dimensional space shown in Figure 1 has been used as the reference.

3.1. Prosody Feature Analysis

The f_0 track was extracted from the JLcorpus using *wrassp* wrapper [26], an advanced speech signal processor library in R computing software [27]. The ksvF0 fundamental frequency estimation function was used with its default settings. The f_0 track was averaged at the sentence level to obtain the mean f_0. Minimum, maximum and range of f_0 were calculated for every sentence. These were then averaged across all sentences to obtain the f_0 statistics for *sad*, *excited*, *neutral* and five secondary emotions, and this result is shown in Figure 2. The dot represents the mean f_0, and the upper and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

lower bounds indicate the maximum and minimum values, respectively. The bold black number at the bottom of the graph is the f_0 range. The plotting was done using R's ggplot [28]. Even though the frequency range for female2 and male2 speakers are different, the effect of emotions on f_0 has common trends. Among the secondary emotions, *enthusiastic* and *anxious* have the highest mean f_0 (high arousal emotions), with *enthusiastic* having the largest range (averaged across male2 and female2 speakers). *Apologetic* (low arousal emotion) has the lowest mean f_0 and range. *Confident* and *worried* fall in between the other three emotions that have more extreme values. Even though emotions like *apologetic*, *confident*, *worried* have similar mean f_0 values, the f_0 range differentiates them. *Confident* in the 4^{th} quadrant of the valence-arousal plane has mean f_0 similar to *neutral*, *worried*; but its range is higher than that of *worried*, which may be an effect of its slightly higher arousal level or positive valence level. Interestingly, the primary emotion *sad* does not have the lowest mean f_0 value despite having the lowest valence level. A more detailed analysis of the fundamental frequency contour of secondary emotions was conducted by the authors and is reported in [29].

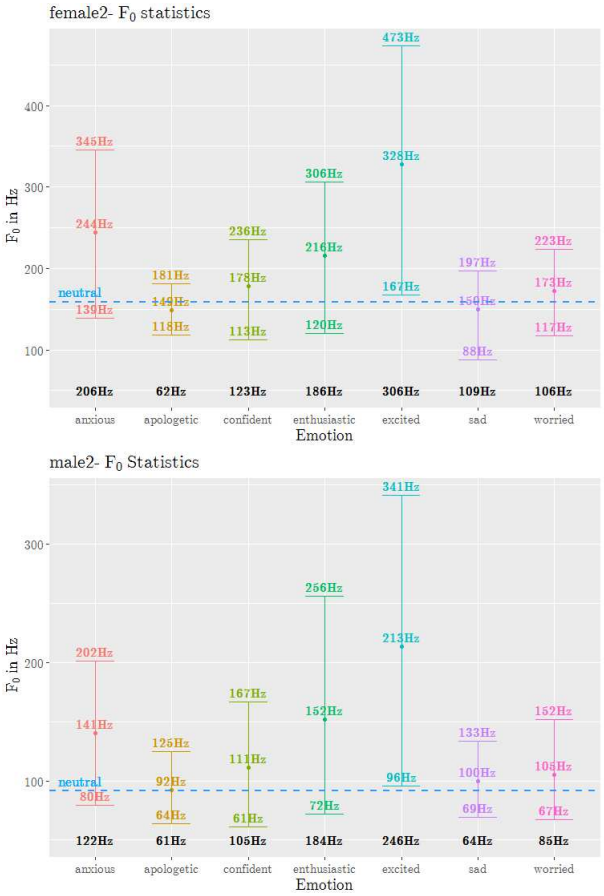

Figure 2. Fundamental frequency statistics of 5 secondary emotions, and *sad*, *excited* and *neutral* from JLCorpus for speakers female2 and male2 speakers. The dot on each line represents the mean f_0, and the upper and lower bounds indicate the maximum and minimum values, respectively. The bold black number at the bottom is the f_0 range. The dash lines represent the results for *neutral*.

Mean intensity in decibels (*dB*) was measured using *wrassp* wrapper using the rmsana short-term root mean square amplitude analysis function with default settings. The intensity was averaged at the sentence level. Figure 3 shows boxplots for intensity across *sad*, *excited*, *neutral*, and 5 secondary emotions plotted using R's ggplot. Each point on the plot represents the mean intensity of a sentence. The dashed line is the mean intensity for *neutral*. It can be seen that emotions influence intensity very strongly. Distinct regions, often with little or no overlap, can be seen in the boxplots. In contrast to the f_0 values, the mean intensity values for the primary emotions *excited* and *sad* are clearly at the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

extremities. Among the secondary emotions, *enthusiastic* and *anxious* (high arousal) have the highest intensity, and *worried* and *apologetic* (low arousal) have the lowest. All high arousal emotions (*anxious*, *enthusiastic*) are much above the *neutral* line and low arousal emotions (*apologetic*, *worried*) are near or below it. *Confident* with arousal levels near *neutral* has intensity values slightly higher than *neutral*. This could be due to the positive valence of *confident*, and this claim needs to be investigated further. Also, *confident* has slightly higher arousal than *worried*, resulting in higher mean intensity values.

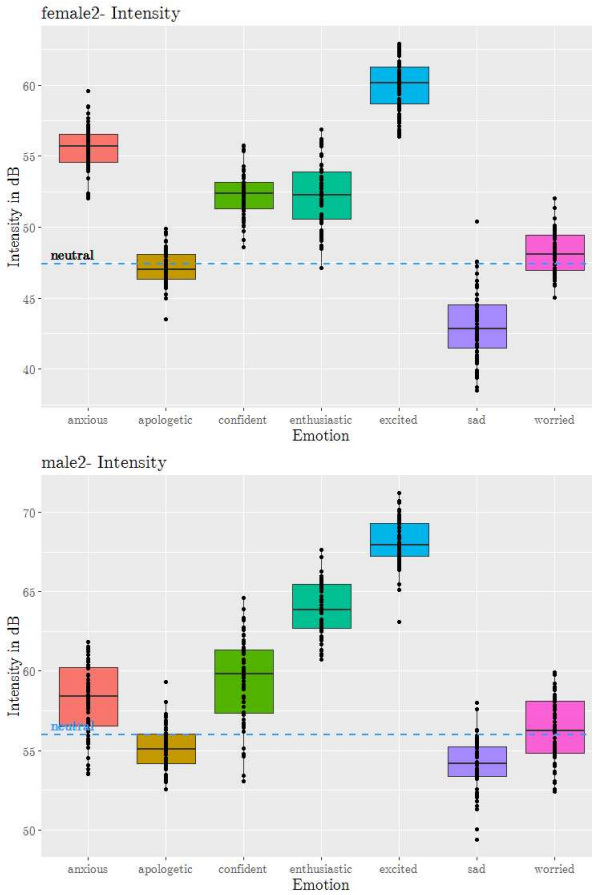

Figure 3. Mean intensity boxplot of 5 secondary emotions, and *sad*, *excited* and *neutral* from JLCorpus for female2 and male2 speakers.

Finally, speech rate in syllables/second was calculated by counting the number of syllables per sentence, and it was divided by the sentence duration [30]. The speech rate variations are not as pronounced as f_0 and intensity due to the short duration of the sentences in the JLCorpus (as noted in [23]). Additional statistical analysis was done to understand the effect of emotions on speech rate. Figure 4 shows the statistical analysis results (from R) for eight emotions, with the value in the box representing the average speech rate for each emotion, where significantly different emotion pairs obtained from a pairwise t-test are marked by arrows. The emotion with the lowest average speech rate is the primary emotion *sad* (low arousal), and is significantly different from all other emotions except *apologetic*, possibly expected due to their similar valence and arousal levels. The primary emotion *excited* (high arousal) has the highest average speech rate, followed by *anxious* and *enthusiastic*. *Confident* shows a significant difference in speech rate from all other emotions. This could be a result of its unique position in the 4^{th} quadrant (See Figure 1). Overall, the results suggest that with reducing levels of arousal from *enthusiastic* to *apologetic*, the speech rate reduces.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

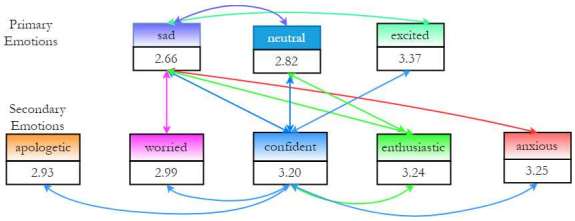

Figure 4. Speech rate statistics of 5 secondary emotions, and *sad*, *excited* and *neutral* from JLCorpus for female2 and male2 speakers.

To summarise, for high arousal emotions (like *anxious, enthusiastic confident*), the feature values for f_0, and mean intensity are high. For low arousal emotions (like *apologetic, sad, worried*), the feature values are low. For speech rate, all emotions follow an increasing trend of feature values as the arousal level of the emotions increase. Thus, all three prosody features are arousal-differentiating. It was found that *confident* behaves similarly to *worried* for the arousal-differentiating features. The results of this analysis suggest that the three prosody features are impacted by secondary emotions. Hence, modelling these three prosody features can be effective in synthesising these secondary emotions.

3.2. f_0 Contour Analysis

Figure 5 shows the time-normalised f_0 contour for five secondary emotions and *neutral* extracted for the same sentence. Comparing *neutral* and secondary emotions, there are clear differences in mean and range of f_0 (also noted in the statistical analysis reported in Section 3.1); most importantly, f_0 contour shape shows considerable differences. For example, in Figure 5, consider the sample points between 10 to 20. One can see that the shape of the contour for *apologetic* is visually different from that of *enthusiastic*, and it is not just a difference in the mean and range values alone; it is in the timing of the peak of the contour. This indicates that even though a qualitative model based on the f_0 statistics can provide emotion separation for primary emotions [5], such a model maybe insufficient to capture the f_0 contour variations of secondary emotions.

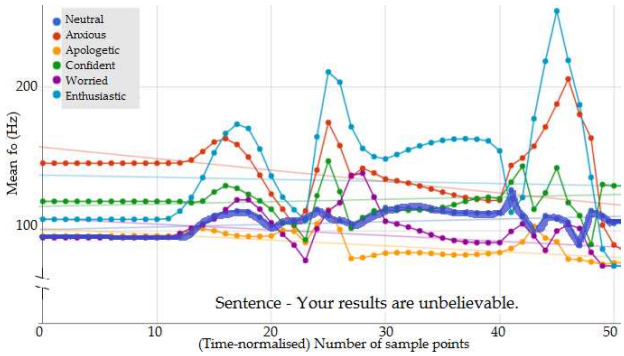

Figure 5. f_0 contour of 5 secondary emotions and *neutral* (dark blue bold line).

Contour-based modelling conveys concurrent linguistic information (e.g. sentence modality, word prominences) and paralinguistic information like emotions [29]. The Fujisaki model is one of the classic f_0 contour models [31]. Fujisaki model approximates the natural f_0 contour and interpolates through unvoiced sounds. The model is event-based, i.e., every command is related to the onset of a new phrase, accented syllable or boundary tone. This model quantifies the f_0 contour with a few parameters using which the f_0 contour can be constructed. Studies reported in [32–36] have used the Fujisaki model for various speech signal applications.

The Fujisaki model [33] parameterises the f_0 contour superimposing (see Figure 6): (1) the base frequency F_b (indicated by the horizontal line at the floor of the f_0 pattern), (2) the phrase component - declining phrasal contours accompanying each prosodic phrase, and (3) the accent component

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

- reflecting fast f_0 movements on accented syllables and boundary tones. These components 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 phrase of the ensuing prosodic phrase. (Phrase command duration *Dur_phr = End of phrase time - T_0*)
2. *Phrase command amplitude (A_p)*: Magnitude of the phrase command that precedes each new prosodic phrase, quantifying the reset of the declining phrase component.
3. *Accent command Amplitude (A_a)*: Amplitude of accent command associated with every pitch accent.
4. *Accent command onset time (T_1) and offset time (T_2)*: The timing of the accent command can be related to the timing of the underlying segments. (Accent command duration *Dur_acc = T_2 - T_1*)

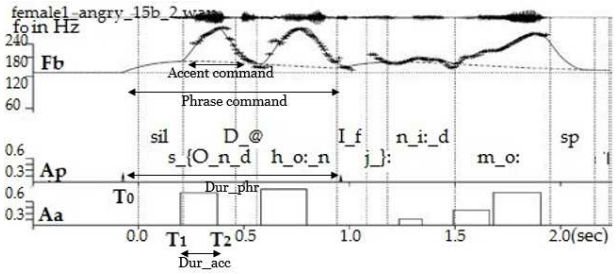

Figure 6. Fujisaki parameters for *'Sound the horn if you need more'* (SAMPA symbols). T_0, T_1, T_2 marked for first phrase and accent commands only.

Using the parameters, the f_0 contour can be obtained as:

$$ln(f_0(t)) = ln(F_b) + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj}(G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j})) \quad (1)$$

where,

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t exp(-\alpha_i t), & \forall t \geq 0. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

$$G_{aj}(t) = \begin{cases} min[\gamma_j, 1 - (1 + \beta_j t) exp(-\beta_j t)], & \forall t \geq 0. \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

F_b - Bias level upon which all the phrase and accent components are superposed to form an f_0 contour
β_j - Natural angular frequency of the j^{th} accent command
α_j - Natural angular frequency of the j^{th} phrase command.
I - Number of phrase commands
J - Number of accent commands
A_{pi} - Magnitude of the i^{th} phrase command
A_{aj} - Magnitude of the j^{th} accent command
T_{0i} - Instant of occurrence of the i^{th} phrase command
T_{1j} - Onset of the j^{th} accent command
T_{2j} - Offset of the j^{th} accent command
γ_j - Ceiling level of the accent component for the j^{th} accent command.

A_a, A_p, T_0, T_1, T_2, F_b, α, β, γ are referred to as the Fujisaki parameters. In this study, α and β are kept constant, and the other six parameters are modelled for the different emotions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

3.2.1. Fujisaki Parameterisation of f_0 Contour

The f_0 contour was extracted with the Praat standard method [37] for every 0.01s. Then the Fujisaki model parameters were estimated from the natural f_0 contour using an automatic algorithm called AutoFuji extractor [38]. In the analysis of reading-style speech, typically, every content word is characterised by at least one accent command associated with the primary pitch accent, and the base frequency F_b is kept constant for each speaker [39]. 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 emphasised can contain two accent commands as seen in the syllable 'm_o:' of Figure 6 (See between time 1.5 to 2 seconds). The Fujisaki model parameters for each utterance were checked to ensure that potential errors in f_0 tracking did not tamper the parameter, leading to additional accent commands in unvoiced segments. This hand-checking was done by the first and fourth authors of this paper, with the first author checking all the files once and correcting them, followed by the fourth author rechecking them. In most cases, the F_b as set for the automatic algorithm was used as it is. But for certain speakers, the F_b has to be adjusted ($\pm$ 10 Hz) as a function of the emotion portrayed to make the Fujisaki-estimated contour better fit the original f_0 contour. This process cannot be conducted automatically because errors in f_0 tracking can cause wrong Fujisaki parameter estimations.

There are 2400 short utterances in the JLCorpus, out of which 1200 were analysed[9]. Only a subset of the corpus was analysed because the Fujisaki parameterisation requires hand-correction. Also, by taking a subset, two renderings of each sentence in the JLCorpus (out of four available) for each emotion for each speaker were analysed. As we aim to synthesise these f_0 contours, getting an accurate parameterisation of the contour for a given sentence is important. Hence, two renderings of the same sentence are used to capture to f_0 contour accurately. The impact of adding more renderings of the same sentence on the model will need to be tested in future when more data gets hand-corrected.

Finally, 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 results are collated, and it contains the Fujisaki model parameters (A_a, A_p, F_b, T_0, T_1, T_2) for each of the syllables. Analysis of the effect of emotions on the Fujisaki parameters (detailed report in [29]) showed that they were affected by the emotions, with accent command parameters (smaller units - A_a and Accent command duration T_2 - T_1) and F_b having the most significant effect.

The analysis in Section 3.1 showed that the secondary emotions could be differentiated using the mean intensity. Hence, a rule-based approach will be used to model the mean intensity of the secondary emotions during synthesis. The intensity contour modelling has not been attempted here, as past research has not well-established an intensity contour model at the accent and phrase level. Hence, developing a new intensity contour model is reserved for future investigation.

A rule-based approach will be used to model the speech rate. Modelling f_0 contour by the Fujisaki model and mean intensity and speech rate using rules addresses research question 1.

4. Emotional TTS Synthesis System Development

Here, we address research question 2. The overall system diagram for the TTS synthesis system is shown in Figure 7. The inputs to the system are the text to be converted to speech and the emotion tag to which the conversion has to be done. Synthesised speech is produced from the input text using a text-to-speech module. This produces synthesised speech that has no emotions. The f_0 contour Fujisaki parameters of the non-emotional speech are extracted via the automatic Fujisaki extractor module. The non-emotional f_0 contour Fujisaki parameters are transformed into emotional f_0 contour parameters.

[9] Both male and female speakers' sentences were used for the initial analysis. But only the male speaker sentences were used for the emotion-based f_0 contour model developed because it is this male speaker's voice that will be synthesised.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

Using these emotional f_0 contour parameters, the f_0 contour corresponding to the emotion tag is reconstructed. The intensity and speech rate decisions are made using the emotion tag. Finally, using the reconstructed f_0 contour, the intensity and speech rate values, emotional speech is resynthesised. Details about each module are given in the following sections.

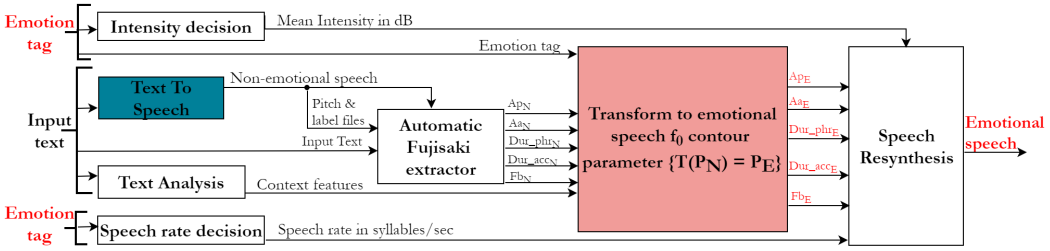

Figure 7. Emotional text-to-speech synthesis system with f_0 contour transformation.

4.1. Text To Speech Module

Previous work [40,41] has led to the development of a TTS synthesis system in New Zealand English based on MaryTTS [42] (with approximately 1000 sentences spoken by a male New Zealand English speaker). Synthesised speech for New Zealand English is currently without any emotion and will be called *non-emotional speech* here. The input text is passed through the New Zealand English MaryTTS system, and the output non-emotional speech is obtained.

4.2. Automatic Fujisaki Extractor

f_0 contour is extracted from the non-emotional speech (by *Praat* Auto Correlation Function method [43]). Label files are obtained from the input text and non-emotional speech using the New Zealand English option of the Munich Automatic Web Segmentation System [44]. The pitch and label files are provided to AutoFuji extractor [38] to obtain the five derived Fujisaki model parameters of non-emotional speech - A_{pN}, A_{aN}, Dur_phr_N, Dur_acc_N, F_{bN}, where N represents "non-emotional". The parameters are then time-aligned to the text at the phonetic level. These Fujisaki parameters are obtained via an automatic process and are not hand-corrected. The Fujisaki parameters will be transformed into corresponding emotional speech parameters. Hence, hand-correction of these non-emotional f_0 contour parameters is not necessary. Avoiding hand-correction also makes real-time synthesis possible, which is suited for human-computer interaction applications.

4.3. Transformation to Emotional Speech Parameters

This module transforms the Fujisaki model parameters of non-emotional speech f_0 contour to that of emotional speech. For conducting this transformation, a regression model was developed, as described here.

4.3.1. Features for the Regression Model

The only inputs available for an emotional TTS synthesis system are the text to be converted to speech and the emotion to which the speech has to be transformed. For real-time implementation, all the features used for transforming the f_0 contour parameters here are based on these two inputs only. A list of all features, along with their extraction methods, is given in Table 3. In total, 109 features are extracted.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

Table 3. Features used for f_0 contour transformation.

Feature	Description	Extraction method
Linguistic context features	Count = 102, Eg. accented/unaccented, vowel/consonant	Text analysis at the phonetic level using MaryTTS.
Non-emotional f_0 contour Fujisaki model parameters	Five Fujisaki parameters - A_{pN}, A_{aN}, Dur_phr_N, Dur_acc_N, F_{bN}	Passing non-emotional speech to AutoFuji extractor.
Emotion tag	Five primary & five secondary emotions	Each emotion tag is assigned to the sentence
Speaker tag	Two male speakers	Speaker tag is assigned

Linguistic context features refer to a set of features that describe the phonetic environment of the target phoneme. In this investigation, the linguistic features used in MaryTTS [42] are used for f_0 contour prediction. This choice is further motivated by the fact that MaryTTS is our front-end synthesiser (more details in Section 4.3.2). Examples of context features are the forward/backward position of a phoneme in a syllable, the number of accented/stressed syllables before/after the current syllable, ToBI end-tone marking etc. This feature extraction process is represented by the *text analysis* module in Figure 7.

The *non-emotional f_0 contour Fujisaki parameters* are used as features for the transformation. The extraction of these features is represented by the *automatic Fujisaki extractor* module in Figure 7 as described above in Section 4.2.

Another feature used is the *emotion tag* representing the emotion to which transformation has to be done. The database for training the transformation model contains two male speakers. Hence, a *speaker tag* is also used as a feature for the transformation model development.

4.3.2. f_0 Contour Transformation Model

The set of hand-corrected Fujisaki parameters (extraction and hand correction described in Section 3.2.1) obtained from the natural emotional speech in JLCorpus is the target value to be predicted by the model. There are 7413 phoneme tokens from two male speakers of JLCorpus. These phoneme tokens and corresponding Fujisaki model parameters form the database for f_0 contour transformation model development. 80% of the database is used for training, and 20% is used for testing using random selection. It was ensured that parts of the same sentences were not split into the train and test set. This was done by choosing three sentences used for testing (which accounts for approximately 20% of the total tokens) and the remaining 12 sentences used for training. (In total, there are 15 different sentences in the JLCorpus, for each emotion, each speaker and each repetition.). As seen in Figure 7, the input to f_0 contour transformation model is non-emotional f_0 contour features, emotion tag and context features. These are the features based on which the f_0 contour transformation model is trained. A less resource-intensive machine learning-based regression model is developed using the hand-crafted features extracted here. The transformation predicts Fujisaki model parameters for every phoneme in an input sentence based on the emotion tag to which the conversion needs to be done.

The f_0 contour transformation model training was as follows.

Here, we choose two ensemble regressors - Random Forest [45], and Adaboost [46] as proof of concept to implement ensemble learning-based regression. Both Random Forest and Adaboost will be allowed to run independently, and the outputs obtained from the two regressors are aggregated in the end without giving any preference to either of the algorithms by taking the mean of the predictions from both algorithms. The implementation was done in Python using the scikit-learn machine learning library RandomForestClassifier, and AdaBoostRegressor packages [47]. Based on 109 features corresponding to each phoneme in the training set, the two regression algorithms are individually trained to learn the patterns of emotional speech f_0 contour parameters. The use of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

ensemble methods has the important advantages of an increase in accuracy and robustness when compared to the use of a single model [48]. This makes ensemble methods suited for applications where small improvements in the predictions have an important impact. This is relevant here, as the requirement is to predict the Fujisaki model parameters accurately, which are numbers, and small variations in them can cause the parameters to change to that of another emotion. The hyperparameters (For *Random Forest* - number of trees, the maximum number of features considered for node splitting, the maximum number of levels in each decision tree, the minimum number of data points placed in a node before the node is split, the minimum number of data points allowed in a leaf node, method of sampling data points and for *Adaboost* - number of estimators, learning rate, number of splits) of these supervised learning methods are tuned via grid search cross-validation, and the best parameter set is used for the training. The mean of the predictions from the two algorithms is taken as the final prediction. Such ensemble learning-based regression models are developed for each emotion. These emotion-dependent models are combined to form the emotion transformation model for the f_0 contour parameters.

4.3.3. Using the Transformation Model

Let each non-emotional speech f_0 contour parameter be called P_N, N stands for 'non-emotional'. Then the developed transformation model ($T(P_N) = P_E$) is applied to this non-emotional speech f_0 contour parameter based on the features. P_E denotes emotional speech parameters. The transformed parameters are then used to produce the f_0 contour of emotional speech. This step is represented by the *Transform to emotional speech f_0 contour parameter* module in Figure 7.

4.4. Speech Rate and Mean Intensity Modelling

Table 4 lists the mean speech rate and mean intensity for each of the five secondary emotions estimated by analysing the JLCorpus (as described in Section 3.1). Based on these intensity and speech rate values, rules were identified for each emotion. These rules were then applied to the speech signal after f_0 contour transformation was performed.

Table 4. Mean speech rate value for five secondary emotions.

Secondary emotion	Mean speech rate (syllables/sec)	Mean intensity (dB)
anxious	3.25	58.24
apologetic	2.93	55.14
confident	3.20	59.50
enthusiastic	3.24	63.91
worried	2.99	56.34

4.5. Resynthesis

The Fujisaki parameters predicted for each phoneme in a sentence are time-aligned to the sentence's phonemes. Accent and phrase commands are placed based on this time alignment, and the F_b is assigned to the sentence. If the model predicts that the accent/phrase command positions need to be changed compared to non-emotional speech, then accent/phrase commands are added/deleted/shifted accordingly. Fujisaki parameters are then used to reconstruct the f_0 contour by superimposing the F_b, accent commands and phrase commands. Once the f_0 contour is reconstructed, emotional speech is re-synthesised by pitch-synchronous overlap-and-add using *Praat*.

5. Performance Analysis and Results

Resynthesised emotional speech was evaluated by a subjective test with 29 participants out of which 14 participants had English (all variants of English included) as their first language (called

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

L1 speakers). 24 participants were from the age group 16-35, and the remaining were distributed over 36-65. All the participants had average, above average or excellent (self-reported) hearing. 23 participants used headphones, 5 used loudspeakers, and the remaining 1 used a laptop speaker. The survey was designed on Qualtrics, a web-based survey platform. The average time taken by the participants to complete the test was 40 minutes. The perception test[10] was divided into five tasks to evaluate the various aspects of the synthesised emotional speech. The participants did the entire test in one sitting. The survey automatically proceeded to the next task when one task was completed. The tasks were presented to the participants in the same order as described here. We also interpret the applicability of these results for a healthcare robot for which this TTS system was developed.

5.1. Task I - Pairwise forced response test for Five Secondary Emotions

This task aims to evaluate if the participants can differentiate five secondary emotions when two of them are presented. Participants listened to 100 sentences and grouped them into the two emotion names provided. These sentences were divided into blocks of 10. Most of the sentences were from the JLCorpus [23]. All five secondary emotions were perceptually evaluated in this pairwise test. This was a forced response test, as the participants could only choose from the emotion list given to them. The 100 sentences were evaluated by the 29 participants, giving 2900 evaluations. The training was provided to the participants to acquaint them with the type of utterances to expect and give them practice in doing the task.

The results of the test are summarised as a collection of confusion matrices, one for each emotion pair shown in Table 5. The horizontal rows indicate the perceived emotions of the participants, and the vertical columns indicate the actual emotions. Each confusion matrix shows the *perception accuracy in percentage* (calculated by the number of correct choices of the emotion divided by the total number of sentences of that emotion as a percentage) for each of the emotion pairs. The highlighted percentage value in the table represents the cases where the participants perceived the actual emotion correctly. The Kappa statistics, $\kappa = 0.816$ (95% Confidence Interval, 0.813 to 0.818, $p < 0.0001$), show strong inter-rater agreement, which means that there was consistency among the participants in differentiating the emotion pairs. Looking at the results, one can deduce that the most confusing emotion pairs were *enthusiastic* vs *anxious*, *confident* vs *enthusiastic* and *apologetic* vs *worried* (which might be due to their closeness in the valence and arousal levels). Each participant evaluated each emotion pair 10 times, ensuring that each emotion pair was evaluated 290 times. So, it is reasonable to assume that the perception accuracy is much higher than chance. Overall, the average perception accuracy across all emotions is 87%. Comparing these results with past studies that have modelled the f_0 contour for emotional speech synthesis such as [49] (reported 50% perception accuracy for expressions good news, bad news and question), [50,51] (reported 75% perception accuracy for *happy*, *angry*, *neutral*) and [52] (reported 65% perception accuracy for *joy*, *sadness*, *anger*, *fear*), the results obtained here are comparable to past works. However, past studies have not reported f_0 contour modelling for these secondary emotions. Hence direct comparison is not possible.

---

[10] This study was approved by the University of Auckland Human Participants Ethics Committee on 24/07/2019 for three years. Ref. No. 023353.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

Table 5. Hit rates from forced response test (ANX: Anxious, APO: Apologetic, CONF: Confident, ENTH: Enthusiastic, WOR: Worried).

	APO	ANX		APO	ENTH
APO	**97.9%**	2.1%	APO	**100%**	0%
ANX	0%	**100%**	ENTH	1.4%	**98.6%**
	CONF	ANX		APO	WOR
CONF	**88.3%**	11.7%	APO	**64.3%**	35.2%
ANX	12.4%	**87.6%**	WOR	32.4%	**67.6%**
	ENTH	ANX		CONF	ENTH
ENTH	**78.6%**	21.4%	CONF	**69%**	31%
ANX	24.8%	**75.2%**	ENTH	30.3%	**69.7%**
	WOR	ANX		CONF	WOR
WOR	**97.9%**	2.1%	CONF	**95.2%**	4.8%
ANX	4.19%	**95.9%**	WOR	22.8%	**77.2%**
	APO	CONF		WOR	ENTH
APO	**94.5%**	5.5%	WOR	**97.9%**	2.1%
CONF	9.7%	**90.3%**	ENTH	0.7%	**99.3%**

In the natural speech subjective test conducted previously by the authors [23], the emotions that were difficult to differentiate were *enthusiastic* vs *anxious*, *confident* vs *enthusiastic* and *worried* vs *apologetic*. From the confusion matrices shown in Table 5, it can be seen that the emotions pairs that are most difficult to differentiate are *worried* vs *apologetic* and *confident* vs *enthusiastic*, as these pairs have the lowest correct hit rates (i.e. the dialog elements of the confusion matrix). The emotions that were most easy to differentiate were *apologetic vs anxious* and *apologetic vs enthusiastic*. These observations indicate that synthesising emotions like *apologetic, worried* may require modelling other acoustic features than the ones considered in this study. From this analysis, it was found that some emotion pairs are easier to differentiate compared to others, and this can be related to the valence-arousal levels of the emotions in the pair. Among the most difficult emotions to differentiate, *apologetic vs worried* and *enthusiastic vs confident* may not be problematic for users of a healthcare robot. This is because, for example, if the healthcare robot is speaking enthusiastically but it is wrongly perceived as confident by the user, it will not negatively impact the user's perception and reaction to the robot. However, the confusion between *enthusiastic vs anxious* will cause difficulty for the users, as a healthcare robot (or any human-computer interaction application) that speaks enthusiastically but is perceived as anxious by the user would baffle the user. Future work on modelling these emotions will have to concentrate on these two emotion pairs in detail. Also, the words in the sentences spoken with these emotions can also help differentiate between *enthusiastic* and *anxious*.

5.2. Task II - Free Response Test for Five Secondary Emotions

In this task, the participants listened to one sentence at a time and wrote down any number of emotions they perceived. This was a free-response test, and the participants were not given any emotion options to choose from. The sentences they heard were a subset of the collection of sentences used for the forced-response tests, and they were different for each emotion. Two sentences corresponding to emotion was evaluated, making a total of 10 sentences evaluated by 29 participants, giving 290 evaluations.

The emotion words written by the participants for each of the five secondary emotions, along with the number of times each word was written, is given in Table 6. It can be seen that the free responses entered by the participants are almost in alignment with the intended emotion. A major confusion was for actual emotion *enthusiastic*, which was reported as *confident* by many participants (24 times). The emotion word *enthusiastic* was also used 21 times for this. Both *enthusiastic* and *confident* have similar valence levels (from Figure 1), which could be the cause of confusion. Also, this confusion between

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

confident and *enthusiastic* may not be detrimental to the experience of the users of a healthcare robot speaking with these emotions.

Table 6. Perceived emotion words in free-response test (Times used).

Actual emotions	Emotion words by participants (count of times used)
Anxious	**Anxious (41)**, Enthusiastic (9), Neutral (4), Confident (3), Energetic (1)
Apologetic	**Apologetic (35)**, Worried (22), Worried/Sad (1)
Confident	**Confident (34)**, Enthusiastic (9), Worried (8), Neutral (5), Authoritative (1), Demanding (1)
Enthusiastic	Confident (24), **Enthusiastic (21)**, Neutral (4), Apologetic (3), Worried (5), Encouraging (1)
Worried	**Worried (38)**, Apologetic (12), Anxious (5), Condescending (1), Confident (1), Neutral (1)

The free-response test was conducted after the forced-response test. The effect of this prior knowledge of emotions names is evident in the responses they have provided. It could be expected that the participants may have given more common emotion words like *happy*, *angry*, *sad* if the free-response test had been conducted before the forced response test. But this ordering was deliberately done to familiarise the participants with the names of the secondary emotions.

5.3. Task III - Naturalness Rating on Five-point Scale

The over-arching extension of this research aims to synthesise emotional voices for healthcare robots. Emotional speech synthesis aims to create emotional voices that are like how humans portray emotions. The naturalness of the emotional voice developed here is evaluated subjectively. In this task, the participants listened to a synthesised emotional sentence and rated the perceived level of naturalness. The question asked to the participants was "Rate the naturalness of this voice (by naturalness, we mean how close this voice is to the human voice) with 5 being the most natural" on a discrete scale of 1 to 5. The 5 levels are based on the levels of naturalness defined in [53], which are very unnatural, unnatural, neutral, natural, and very natural. The sentences used were a subset of the sentences used for the forced-response test in Task I. For each secondary emotion, two sentences were evaluated by 29 participants, giving a total of 290 evaluations.

Figure 8 shows the mean opinion score of naturalness. The average score across all five emotions is also included. The perception of *enthusiastic* was found to be the least natural. However, all emotions' naturalness rating is greater than 2.5, which indicates that the voice was not perceived as unnatural (based on the five naturalness levels described in [53]). The fact that even though the emotions are modelled on *synthesised speech*, the participants still felt that the emotional sentences are close to having a natural quality is a positive result.

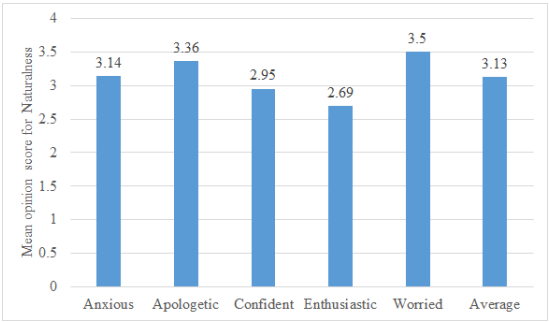

Figure 8. Naturalness rating boxplots for synthesised secondary emotional speech in percentage.

5.4. Task IV - Comfort Level Rating on Five-point Scale

As emotional speech developed here is for healthcare robots, the people listening to the speech from the robots have to find it comfortable to listen to. In this task, the participants listened to a

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

synthesised emotional sentence and rated their perceived comfort level. The question asked to the participants was - "Rate your comfort level in listening to this voice for a long time (By comfort level, we mean if you can listen to this voice for a long time - more than 1 minute) with 5 being most comfortable". Comfort level is the ease of listening, as defined by [53] as the ease of listening to the voice for long periods of time. The participants had to rate the voice on a five-point discrete scale, with 5 being the most comfortable. These five levels are based on the levels defined by [53] ranging from very difficult, difficult, neutral, easy and very easy, as the discrete levels vary from 1 to 5. For each secondary emotion, one sentence was evaluated by 29 participants, giving a total of 145 evaluations. The sentences used for this section are a subset of the collection of sentences used in Task I.

Figure 9 shows the mean opinion score of the comfort level perception. The average score across all five emotions is also included. All the emotions have a mean opinion score above 2 on the comfort level five-point scale, which indicates neutral, easy or very easy to listen to [53]. The perception of *anxious* was found to have the least comfort rating among all five emotions evaluated. For the emotional text-to-speech synthesis developed here, the voice synthesised here is expected to be suitable as, on average, participants found it comfortable to listen to (2.97 mean opinion score).

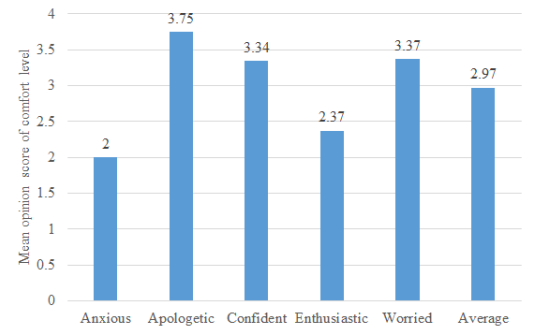

Figure 9. Comfort level rating boxplots for synthesised secondary emotional speech in percentage.

6. Discussion

In this paper, the focus was on developing a low-resource approach for emotional speech synthesis of five secondary emotions by modelling prosody features. The three prosody features have been well-studied and used for the synthesis of primary emotions and speaking styles [10,13,15,20]. Here, we have analysed the impact of secondary emotions on these three prosody features. The acoustic analysis results show that these three prosody features are impacted by secondary emotions. This was the motivation to develop a proof-of-concept emotional TTS synthesis system that synthesises secondary emotions by modelling the three prosody features only. Subjective tests provide a strong indication that the approach of modelling these prosody features is promising for secondary emotions synthesis, at least for the five secondary emotions considered here. This justifies future work in which we will pursue a deeper investigation of incorporating other acoustic features for synthesis. Such an investigation is essential for the emotion pairs that have been found to be the most confusing in the perception tests like *apologetic-worried.*

Instead of using the qualitative statistics of f_0, we have focused on a f_0 contour modelling. Contour-level modelling of f_0 was used for emotional speech synthesis in a study [15] based on the ToBI model (a qualitative model). We expect the subtleties in the f_0 contour can be picked up by quantitative modelling rather than a qualitative approach. Also, by using the ToBI model, one can only get tags for the various tones and breaks, and then the f_0 values will have to be calculated by other approaches. The Fujisaki model, on the other hand, will provide an equation for the f_0 contour, thereby facilitating re-synthesis. Also, this approach seems to be picking up the subtle changes in the f_0 contour introduced by the secondary emotions (as seen in the results in Section 3.2). Hence, in this investigation, we utilise the traditional hand-crafted feature extraction approach and marry it with modern machine learning to effectively synthesise secondary emotions. There is an emphasis on

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 January 2023 doi:10.20944/preprints202301.0008.v1

how the f_0 contour data was prepared for the modelling. This involved parameterising the f_0 contour using the Fujisaki model and hand-correction - a task requiring phonetic knowledge about accents and phrases.

Most emotional synthesisers use quantitative models like mean or range variations on prosody features to synthesise emotional speech [13,20]. Or they rely on spectrogram-based features and use deep learning models to learn patterns for speech synthesis [11,18]. The former statistical features-based model can miss the subtleties introduced in the f_0 contour by the secondary emotions. And the latter relies on a large database, which is not easy to develop for all emotions and all languages. The approach presented in this study utilises the strength of speech signal processing models to develop a low-resource emotional speech synthesiser. This approach can be easily trained for other emotions as well, even if a small database of the emotion is available.

The secondary emotions studied here are novel to emotional speech synthesis research, even though they are commonly used in human conversations. It was found that all three prosody features are arousal-differentiating features. But, valence-differentiating features may be particularly crucial for the secondary emotions, as the arousal level differences for these nuanced emotions are not as dominant as the primary emotions. For e.g. for secondary emotions, the arousal difference between *worried*, *confident*, *enthusiastic* is not much, and to differentiate them, valence-level features may also be needed. This will be a focus of future research. The subjective test results show that at least three out of the five emotions can be adequately modelled by this approach. This can be seen in Table 5 confusion matrix and Table 6, where emotions *anxious*, *enthusiastic* and *confident* could be perceived well be participants. However, emotions *apologetic* and *worried* seem to be confused with one another and are not very well recognised in the free response test as well. These emotions will have to be studied in detail to model them better.

The parametric modelling of the f_0 contour was a crucial addition. This contributes to a low-resource approach to emotional speech synthesis. This can be further expanded by collecting small databases for other secondary emotions and creating models for them. New studies have used sequence-to-sequence modelling [52] for predicting the f_0 contour. Rather than a direct prediction of the contour, a prediction of the f_0 contour Fujisaki model parameters using sequence-to-sequence framework may be a better approach that facilities re-synthesis. But such a neural network-based approach may require a larger database, and further experimentation needs to be done on the feasibility of the approach. Speech research initially tried to model the human speech production system. With emerging trends to employ deep learning in speech technology research, all the features are extracted by automatic processes and fed into a “black box", thus often lacking an understanding of the acoustics features impacted by emotions. This research, on the other hand attempts to understand three prosody features and their impact on secondary emotions, which have been used to synthesise five secondary emotions.

7. Conclusion

This paper reports the addressing of two research questions. One was to understand if prosody features can be used to model secondary emotions. It was found that f_0 contour, speech rate, and mean intensity are impacted by the five secondary emotions. Based on statistical analysis, it was decided to model the f_0 contour by using a quantitative model called the Fujisaki model, and the other two features by rules. The second research question addressed was to develop a TTS synthesis system for secondary emotions. To address this question, a transformation model was developed to transform the f_0 contour Fujisaki parameters to that of emotional speech. The features used for the transformation are the input text and the emotions tag. Once the transformation of the f_0 contour was done, the speech signal was re-synthesised to produce the intended emotion. Also, a detailed subjective test was conducted to evaluate the performance of the emotional speech synthesis system.

What makes this study different from past studies is the attempt to synthesise less explored secondary emotions. Modelling the f_0 contour quantitatively instead of only using qualitative measures

could capture the subtle changes in the f_0 contour due to the secondary emotions, and also it facilitates direct re-synthesis. While modelling the stronger primary emotions, the mean and range features may be sufficient. However, for the secondary emotions, changes in the accent and phrase levels have to be captured.

In future, an emotion-based linguistic analysis of the input text will be advantageous for more accurate predictions of accent and phrase commands. Development of larger datasets with variations like emotions, speakers and linguistic contexts can produce a more robust emotion transformation model, which may be expensive and difficult to obtain. The focus on understanding the properties of the speech signal and what features differentiate one emotion from another is the approach followed here. This has proved to be beneficial in cases similar to this study, where the database is not large enough to perform advanced machine learning-based modelling. Only three prosody features are modelled here. This modelling has produced above-chance results in correctly identifying the secondary emotions. Future research can look at how the other acoustic features, such as spectral and glottal features, can be modelled and incorporated into an emotional TTS synthesis system for secondary emotions.

Author Contributions: Conceptualisation, J.J; Methodology, J.J, B.T.B, C.W. and H.M; Result analysis: J.J, B.T.B; Perception Test Design and analysis: J.J, C.W., Writing - original draft preparation, J.J.; writing - review and editing - J.J., B.T.B, C.W., H.M., All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The authors would like to thank the participants of the perception test for their time and effort.

References

1. Eyssel, F.; Ruiter, L.D.; Kuchenbrandt, D.; Bobinger, S.; Hegel, F. 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. Int. Conf. on Human-Robot Interaction, USA, 2012, pp. 125–126.
2. James, J.; Watson, C.I.; MacDonald, B. Artificial empathy in social robots: An analysis of emotions in speech. IEEE Int. Symposium on Robot & Human Interactive Communication, China, 2018, pp. 632–637.
3. James, J.; Balamurali, B.; Watson, C.I.; MacDonald, B. Empathetic Speech Synthesis and Testing for Healthcare Robots. *International Journal of Social Robotics* **2020**, pp. 1–19.
4. Ekman, P. An argument for basic emotions. *Cognition & emotion* **1992**, *6*, 169–200.
5. Schröder, M. Emotional Speech Synthesis: A Review. Eurospeech, Scandinavia, 2001, pp. 561–64.
6. Becker-Asano, C.; Wachsmuth, I. Affect Simulation with Primary and Secondary Emotions. Intelligent Virtual Agents. IVA 2008. LNCS. Springer, 2008, Vol. 5208, pp. 15–28.
7. Damasio., A., Descartes' Error, Emotion Reason and the Human Brain); Avon books, New York, 1994.
8. James, J.; Watson, C.; Stoakes, H. Influence of Prosodic features and semantics on secondary emotion production and perception. Int. Congress of Phonetic Sciences, Australia, 2019, pp. 1779–1782.
9. Murray, I.R.; Arnott, J.L. Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication* **1995**, *16*, 369–390.
10. Tao, J.; Kang, Y.; Li, A. Prosody conversion from neutral speech to emotional speech. *IEEE transactions on Audio, Speech, and Language processing* **2006**, *14*, 1145–1154.
11. Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.J.; Clark, R.; Saurous, R.A. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. Int. Conf, on Machine Learning, Sweden, 2018, pp. 4693–4702.
12. An, S.; Ling, Z.; Dai, L. Emotional statistical parametric speech synthesis using LSTM-RNNs. APSIPA Conference, 2017, pp. 1613–1616.
13. Vroomen, J.; Collier, R.; Mozziconacci, S. Duration and intonation in emotional speech. Third European Conference on Speech Communication and Technology, Germany, 1993.
14. Masuko, T.; Kobayashi, T.; Miyanaga, K. A style control technique for HMM-based speech synthesis. International Conference on Spoken Language Processing, Korea, 2004.

15. Pitrelli, J.F.; Bakis, R.; Eide, E.M.; Fernandez, R.; Hamza, W.; Picheny, M.A. The IBM expressive text-to-speech synthesis system for American English. *IEEE Transactions on Audio, Speech, and Language Processing* **2006**, *14*, 1099–1108.
16. Yamagishi, J.; Kobayashi, T.; Tachibana, M.; Ogata, K.; Nakano, Y. Model adaptation approach to speech synthesis with diverse voices and styles. IEEE International Conference on Acoustics, Speech and Signal Processing, USA, 2007, pp. IV–1233.
17. Barra-Chicote, R.; Yamagishi, J.; King, S.; Montero, J.M.; Macias-Guarasa, J. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech communication* **2010**, *52*, 394–404.
18. Tits, N. A Methodology for Controlling the Emotional Expressiveness in Synthetic Speech-a Deep Learning approach. International Conference on Affective Computing and Intelligent Interaction, UK, 2019, pp. 1–5.
19. Zhang, M.; Tao, J.; Jia, H.; Wang, X. Improving HMM based speech synthesis by reducing over-smoothing problems. Int, Symposium on Chinese Spoken Language Processing, 2008, pp. 1–4.
20. Murray, I.; Arnott, J.L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of Acoustic Society of America* **1993**, *93(2)*, 1097–1108.
21. Masuko, T.; Kobayashi, T.; Miyanaga, K. A style control technique for HMM-based speech synthesis. International Conference on Spoken Language Processing, Korea, 2004, pp. 1437–1440.
22. Yamagishi, J.; Kobayashi, T.; Tachibana, M.; Ogata, K.; Nakano, Y. Model adaptation approach to speech synthesis with diverse voices and styles. International Conference on Acoustics, Speech and Signal Processing, USA, 2007, p. 1236.
23. James, J.; Tian, L.; Watson, C. An open source emotional speech corpus for human robot interaction applications. Interspeech, India, 2018, pp. 2768–2772.
24. Scherer, K. What are emotions? And how can they be measured? *Social Science Information* **2005**, *44*, 695–729.
25. Paltoglou, G.; Thelwall, M. Seeing Stars of Valence and Arousal in Blog Posts. *IEEE Trans. of Affective Computing* **2013**, *4(1)*, 116–23.
26. Winkelmann, R; Harrington, J.J.K. EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* **2017**, *45*, 392 – 410.
27. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.
28. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer New York, 2016.
29. James, J.; Mixdorff, H.; Watson, C. Quantitative model-based analysis of F_0 contours of emotional speech. International Congress of Phonetic Sciences, Australia, 2019, pp. 72–76.
30. Hui, C.T.J.; Chin, T.J.; Watson, C. Automatic detection of speech truncation and speech rate. SST, New Zealand, 2014, pp. 150–153.
31. Hirose, K.; Fujisaki, H.; Yamaguchi, M. Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1984, Vol. 9, pp. 597–600.
32. Nguyen, D.T.; Luong, M.C.; Vu, B.K.; Mixdorff, H.; Ngo, H.H. Fujisaki Model based f0 contours in Vietnamese TTS. Int. Conf. on Spoken Language Processing, Korea, 2004, pp. 1429–1432.
33. Mixdorff, H.; Cossio-Mercado, C.; Hönemann, A.; Gurlekian, J.; Evin, D.; Torres, H. Acoustic correlates of perceived syllable prominence in German. Annual Conference of the International Speech Communication Association, Germany, 2015, pp. 51–55.
34. Gu, W.; Lee, T. Quantitative analysis of f0 contours of emotional speech of Mandarin. ISCA Speech Synth. Wksp/, 2007, pp. 228–233.
35. Amir, N.; Mixdorff, H.; Amir, O.; Rochman, D.; Diamond, G.M.; Pfitzinger, H.R.; Levi-Isserlish, T.; Abramson, S. Unresolved anger: Prosodic analysis and classification of speech from a therapeutic setting. Speech Prosody, USA, 2010, p. 824.
36. Mixdorff, H.; Cossio-Mercado, C.; Hönemann, A.; Gurlekian, J.; Evin, D.; Torres, H. Acoustic correlates of perceived syllable prominence in German. Annual Conference of the International Speech Communication Association, Germany, 2015, pp. 51–55.
37. Boersma, P.; Weenink, D. Praat: Doing phonetics by computer [Computer program]. Version 6.0.46, 2019.
38. Mixdorff, H. A novel approach to the fully automatic extraction of Fujisaki model parameters. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Turkey, 2000, pp. 1281–1284.

39. Mixdorff, H.; Fujisaki, H. A quantitative description of German prosody offering symbolic labels as a by-product. International Conference on Spoken Language Processing, China, 2000, pp. 98–101.
40. Watson, C.I.; Marchi, A. Resources created for building New Zealand English voices,. Australasian International Conference of Speech Science and Technology, New Zealand, 2014, pp. 92–95.
41. Jain, S. Towards the Creation of Customised Synthetic Voices using Hidden Markov Models on a Healthcare Robot. Master's thesis, The University of Auckland, New Zealand, 2015.
42. Schröder, M.; Trouvain, J. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology* **2003**, *6*, 365–377.
43. Boersma, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Institute of Phonetic Sciences* **1993**, *17*, 97–110.
44. Kisler, T.; Schiel, F.; Sloetjes, H. Signal processing via web services: the use case WebMAUS. Digital Humanities Conf., 2012, pp. 30–34.
45. Liaw, A.; Wiener, M. Classification and Regression by Random Forest. *R news 2.3* **2002**, *23*, 18–22.
46. Yoav, F.; Robert E, S. Experiments with a new boosting algorithm. Int. Conf. on Machine Learning, Italy, 1996, pp. 148–156.
47. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
48. Mendes-Moreira, J.; Soares, C.; Jorge, A.M.; Sousa, J.F.D. Ensemble approaches for regression: A survey. *Acm computing surveys (csur)* **2012**, *45*, 1–40.
49. Eide, E.; Aaron, A.; Bakis, R.; Hamza, W.; Picheny, M.; Pitrelli, J. A corpus-based approach to expressive speech synthesis. ISCA ITRW on speech synthesis, USA, 2004, pp. 79–84.
50. Ming, H.; Huang, D.Y.; Dong, M.; Li, H.; Xie, L.; Zhang, S. Fundamental Frequency Modeling Using Wavelets for Emotional Voice Conversion. International Conference on Affective Computing and Intelligent Interaction, China, 2015, pp. 804–809.
51. Lu, X.; Pan, T. Research On Prosody Conversion of Affective Speech Based on LIBSVM and PAD Three Dimensional Emotion Model. Wkhp on Advanced Research & Tech. in Industry Applications, 2016, pp. 1–7.
52. Robinson, C.; Obin, N.; Roebel, A. Sequence-To-Sequence Modelling of F_0 for Speech Emotion Conversion. International Conference on Acoustics, Speech, and Signal Processing, UK, 2019, pp. 6830–6834.
53. Viswanathan, M.; Viswanathan, M. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language* **2005**, *19*, 55–83.