APPLICATION OF FUJISAKI MODEL TO DERIVE FEATURES TO CAPTURE PROSODIC INFORMATION.

Mr.Amrik Singh[1], Mr. R.K. Sharma[2]
[1,2]Guru Kashi University, Talwandi Sabo

ABSTRACT

The usefulness of a model for characterising pitch profiles in voice signals is critical in a wide range of application areas, but it is particularly important in natural-sounding text-to-speech systems, which are becoming increasingly popular. Despite its simplicity, the Fujisaki model has demonstrated remarkable accuracy across a wide range of languages. A much more difficult task is that of solving the inverse problem, i.e., extracting the input parameters that formed an observed pitch contour, which has the potential to be very beneficial in the field of automatic extraction of prosodic parameters from a given speech signal and could be of considerable importance A tiny sample of 100 male and female utterances from the natural, USS, and HTS systems were used to establish the Fujisaki Model parameters speakers was used. Both natural and synthetic speech are produced using the same text content.

KEYWORDS: Utterances, Model, Contour, Prosodic, Speech.

I. INTRODUCTION

Computational models are increasingly being used to capture prosodic characteristics. A good example of this is when attempting to parameterize fundamental frequency (F0) contours sparingly, that is, by establishing By extracting a limited number of parameters from the estimated F0 contour values, we may establish links between the F0 contour and the utterance's informative units and structures. An algorithm based on data from an error-prone method that relies on F0 is inherently sensitive to those flaws because F0 only exists for spoken sounds and proper extraction is frequently difficult. Despite the presence of non-vocalic pauses and micro-prosodic undulations in F0,the auditory system interprets these as truly seamless intonation contours. Because the laryngeal mechanism is so difficult to describe in terms of articulatory features, the fundamental frequency contour cannot be described in terms of how it is produced. This mechanism may be shared by all people, although previous studies have found that speakers of different languages use intonation differently and some are more sensitive to specific parts of the F0 contour than others. When speaking a stress-timed language, such as English or German, speakers pay close attention to accented syllables and boundary tones, whereas tone language speakers give each word its own tonal contour.

FUJISAKI'S MODEL

Between the 1970s and the 1980s, H. Fujisaki and his colleagues devised an analytical model for characterising the fundamental frequency (F0) changes in the electromagnetic spectrum. It

encompasses the basic mechanisms that are involved in the creation of speech and are responsible for the formation of a specific prosodic structure. Following that, the representation of speech prosody in terms of the model characteristics, also known as the inverse problem, has been handled using a variety of approaches and techniques. The overlapping between sets of model contours and natural contours has been demonstrated to be satisfactory using Fujisaki's model (see Fig. 1).

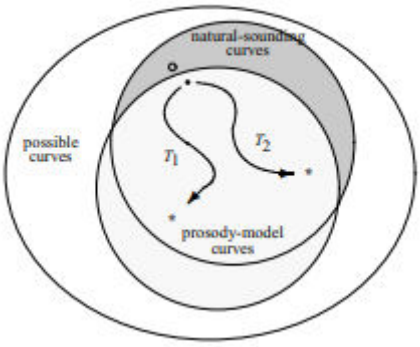

Figure 1: Using a prosody method to create speech synthesis, the following steps are performed: pitch contour extraction from speech (◦), pitch contour representation by using a prosody model (•), and possible modified pitch contours generated by using a prosody model (*) with appropriate prosodic-feature manipulations (Ti).

II. RESEARCH METHODOLOGY

First, the Fujisaki Model parameters were examined using a small sample of 100 natural, USS, and HTS utterances provided by a man and female speaker, respectively, in order to determine their significance. Both natural and synthetic speech are produced using the same text content. In this study, the parameters derived from the model-generated F0 contour in log-domain were used in the analysis; these were Fb, xp, yp, xa, and ya, to name a few. Shows the summary of the analysis in based on mean and standard deviation, as well as the number of observations. On the basis of the same text, we conduct the current analysis on individual utterances, and we then expand the application of the application of the technique to a classification problem on a large, non-parallel, statistically significant dataset

DATA ANALYSIS

Due to the varied durations of the commands and elements, it is not possible to use the Fujisaki model parameters as strength. Our early work utilising the characteristics of the Fujisaki model for judging natural and synthetic speech in Gujarati are investigated. Figure 2 shows spectrograms for the same phrase for human speech, USS, and HTS-based synthetic speech. In terms of speaker characteristics, the spectrogram of USS-based synthetic speech is very comparable to the spectrogram of authentic speech. Although the spectrogram has breaks that appear to indicate a discontinuity in the formant contour, this is not the case

(dotted oval showing abruptness due to concatenation). These gaps can occur in actual speech as well; however, because to the concatenation of speech sound units, the breaks are longer. frequency with which they occur in USS-based speech is significantly higher than in normal speech. The appears to be the formant structure and intelligibility of HTS-based speech are not preserved in HTS-based speech, as evidenced by the spectrogram (dotted squares).

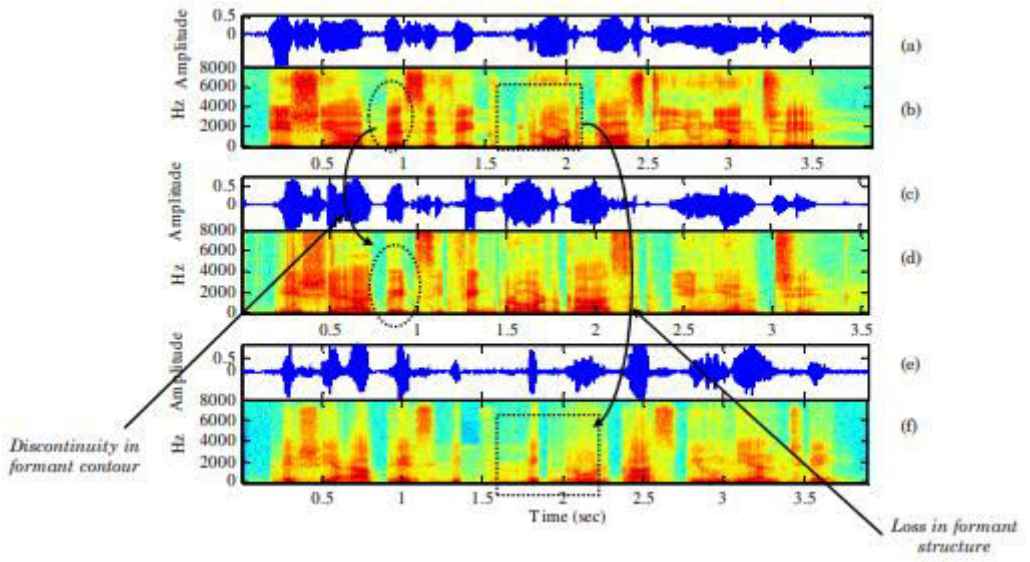

Figure 2 (a) Speech Signal, (b) spectrogram of (a), (c) USS-based speech, (d) spectrogram of (c), (e) HTS-based speech and (f) spectrogram of (e).

With the Itakura–Saito distance metric, we are able to gauge the differences between various spectrograms. A perceptual disparity among an original spectrum P(ω) and an approximation spectrum P(ω) is measured by this function. A synthetic representation of natural speech is considered should be a close match to real speech To compute the IS distance, the utterances were time-aligned using DTW, and the LPCs were extracted from the speech signal for every 20 ms speech frame with a frame shift of 10 ms. between the utterances and the target. background noise.

$$D_{IS}(P(\omega),\hat{P}(\omega))=\frac{1}{N}\sum_{m=1}^{N}\left[\frac{P(\omega_m)}{\hat{P}(\omega_m)}-\log\left(\frac{P(\omega_m)}{\hat{P}(\omega_m)}\right)-1\right],$$

The number of speech frames is denoted by the letter N. Table 1 shows that the IS distance between natural and USS speech is much shorter when compared to HTS speech, which is owing to the fact that the IS distance evaluates spectral properties that are reliant on the size and form of the person's vocal tract.

Table 1: Over 100 utterances, the average IS distance between natural and synthetic speech was calculated for both male and female speakers.

D_{IS}	USS		HTS	
	Male	Female	Male	Female
IS	11.675754	9.2565341	14.994685	18.448603

4.1

Statistical Evaluation of Outcomes

Figure 3 depicts the scatter plots for 100 USS and 100 HTS synthesised utterances formed by the mean of accent and sentence elements, for each group. The size and form of the USS and natural speech clusters differ from those of the HTS and natural speech groups, respectively. It is discovered, in particular, that the clusters for USS synthetic speech and natural speech are more overlapping, making differentiation between the two classes of speech more difficult. However, when it comes to HTS speaking (especially female voice), the two types can be identified more clearly. As a result, the female voice in HTS lacks many of the prosodic features found in the natural voice. The Student's t-test was used to see if the parameter distributions for natural and synthetic voices differed significantly from one another. We also investigated if the two sets of synthetic voices differed considerably from one another. This demonstrates how the phrase and accent parameters might be useful. to be an effective collection of criteria for distinguishing between genuine and synthetic speech in the future.

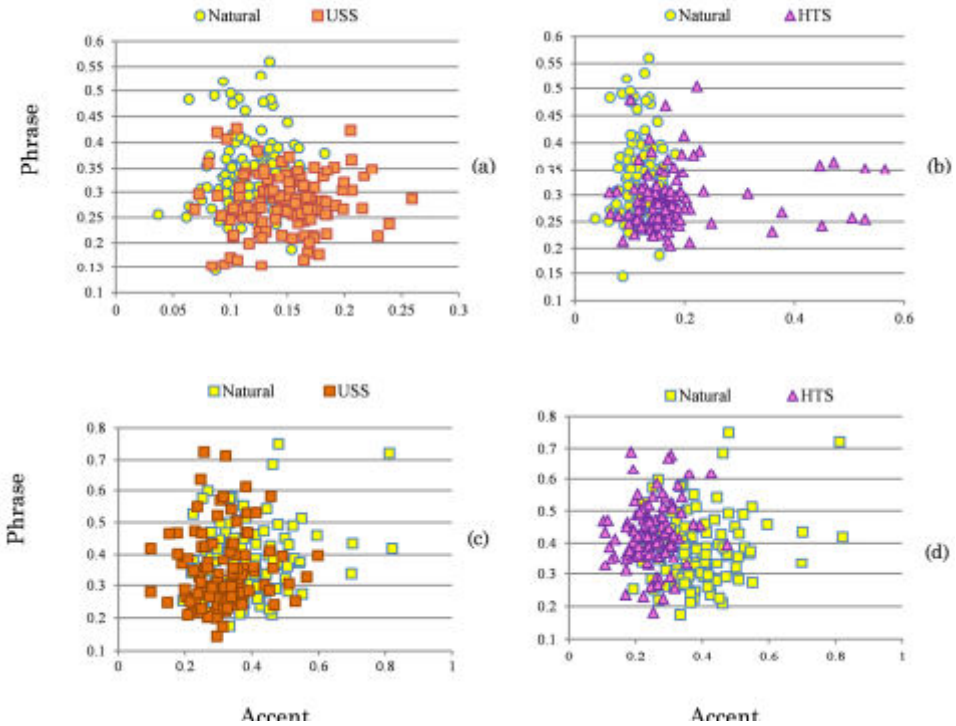

Figure 3: Groups of accent and phrase components for (a) natural vs. USS (male), (b) natural vs. HTS (male), (c) natural vs. USS (female) and (d) natural vs. HTS (female)

III. EXPERIMENTAL OUTCOMES

Creating features using the operation of Fujisaki model parameters is complex because information about prosodic phrase breaks and accent components is incorporated in the place of their occurrence and is only accessible when the utterances are given parallel. As a result, generalising these features to non-parallel utterances is not always straightforward and can be difficult. In this Section, we'll create a feature vector that incorporates both the F0 and the predicted sentence and accent variables., among other things. The ASV spoof challenge database is used to assess the functionality of this feature set.

5.1 Outcomes on the Development Set of ASVspoof challenge Database

Because of the small incorporate perspective, we propose a hidden preliminary to evaluate the percent EER of Fujisaki model-based features for different Gaussian blend parts using the Fujisaki model. The number of components in the combination might range from 2, 4, 8, 16, 32, 64, and 128. When the amount of mixture components is increased, it is noted that the overall tendency is a drop in the EER value. In Figure 4, the lowest EER of 40.88 percent is found for 128 combination parts, which is the lowest in the study. As a result, we use 128 mixture components in this work, as we have in all of our past tests, in order to ensure homogeneity.

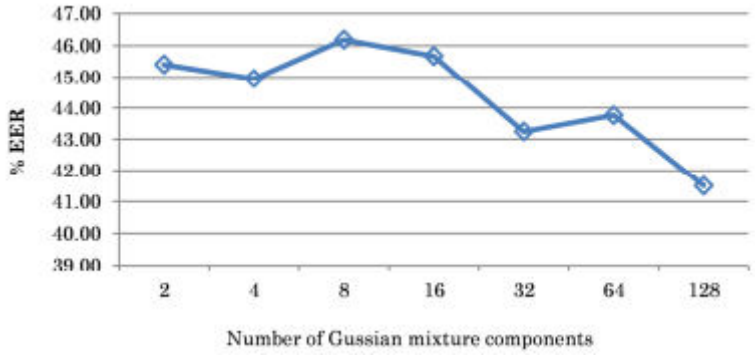

Figure 4: The percent EER achieved on the development set for the Fujisaki model-based features with a changing number of Gaussian constituent materials was calculated for each feature on the development set.

In the following section, we demonstrate the effectiveness of the Fujisaki model-put together highlights with respect to the improvement set of the ASV parody challenge data set (see Figure 1). As displayed in Table 2, the normal percent EER created by joining 128 blend parts is very high, averaging around 40.88 percent overall. It ought to be underlined that the improvement set contains only of vocoder-based spoofs, and as needs be, the percent EER

gained is extremely high in the current situation. A score-level blend of the Fujisaki model features and the structure based MFCC, CFCCIFS, and SBAE highlights is additionally endeavoured to reveal any possibly supplementing data. Be that as it may, as displayed in Table 5.18, in any event, when score-level combination is performed with any weight factor, the framework based highlights show no genuinely critical increment over the discoveries gained with the F0, SoE1, SoE2, and forecast based highlights (i.e., no huge improvement). As a result, in order to make this proposed feature vector usable for the SSD task, it must be adjusted in an efficient manner.

Table 2: Score-level fusion of Fujisaki model-based feature set with system-based feature sets (using D3 feature vector) at various fusion factors f on the development set yielded the following EER (in percent)

Feature Set1	**FusionFactor (αf)**										**Feature Set2**
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
Fujisaki	40.88	15.99	80.002	4.500	3.200	2.500	2.123	1.900	1.630	1.700	1.800	**MFCC**
	40.88	15.01	7.412	4.200	3.000	2.200	1.800	1.700	1.487	1.600	1.700	**CFCC**
Features	40.88	19.99	9.210	4.600	3.000	2.000	1.600	1.412	1.258	1.199	1.230	**CFCCIFS**
	40.88	17.300	9.700	5.633	4.000	2.740	2.002	1.900	1.544	1.490	1.490	**SBAE**

IV. CONCLUSION

It was the first time that the Fujisaki model was investigated for the purpose of identifying the absence of prosodic data in discourse, rather than its ordinary application in prosody change for TTS frameworks. Thinking about the way that the ridiculed discourse rejects the prosodic characteristics that are available in veritable discourse, we endeavor to sum up our discoveries from the expression and complement parts to non-equal expressions. While looking at the F0, SoE1, and SoE2 highlights with their elements, the sole consistency was noticed in the fact that the percent EER decreased as the amount of dynamic information increased for each of the three features. The use of a prosodic model while generating speech synthesis makes it more difficult for these criteria to distinguish between spoof-specific aspects. As a result, it is necessary to change and construct highlights that perform

fundamentally well on the ASV parody challenge data base, and following that to investigate whether these features can be generalised.

REFERENCES

1. H. Fujisaki, "Information, prosody, and modelling," in Proceedings of Speech Prosody, Nara, Japan, March 2004, pp. 1-10.

2. H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations," Dept. for Speech, Music and Hearing, KTH Stockholm, Quarterly Progress and Status Report, 1981.

3. H. Fujisaki, S. Ohno, and W. Gu, "Physiological and physical mechanisms for fundamental frequency control in some tone languages and a command–response model for generation of their F0 contours," in Int. Sym. on Tonal Aspects of Lang.—with Emphasis on Tone Lang., Beijing, China, 2004, pp. 61-64.

4. A Rajpal, et al., "Native language identification using spectral and source-based features," in Int. Speech Comm. Assoc. (INTERSPEECH), San Francisco, USA, 2016, pp. 2383-2387.

5. Sakurai and H. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," in 4th Int. Conf. on Spoken Lang. Process., (ICSLP), vol. 2, Philadelphia, PA, 1996, pp. 817-820.

6. S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," in IEEE Int. Conf. Acoust., Speech, Signal Process., (ICASSP), 2002, pp. 509-512.

7. T. B. Patel and H. A. Patil, "Analysis of natural and synthetic speech using Fujisaki model," in IEEE Int. Conf. Acous., Speech and Sig. Process. (ICASSP), Shanghai, China, 2016, pp. 5250-5254.

8. A E. Aronson and D. M. Bless, Clinical Voice Disorders. New York: Thieme Medical Publishers, 2009.

9. V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted," Jour. Acoust. Soc. Amer. (JASA), vol. 133, no. 15, pp. 3050-3061, May 2013.

10. T. Patel and H. Patil, "Novel approach for estimating length of the vocal folds using Fujisaki model," in Int. Symposium on Chinese Spoken Lang. Process. (ISCSLP), Singapore, 2012, pp. 308-312.