On the Use of Fujisaki Parameters for the Quality Prediction of Synthetic Speech

Florian Hinterleitner[1], *Christoph Norrenbrock*[2], *Sebastian Möller*[1]

[1]*Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany*
[2]*Digital Signal Processing and System Theory, CAU Kiel, Germany*
florian.hinterleitner@telekom.de

Abstract: This paper presents research on the use of Fujisaki parameters for the quality prediction of synthetic speech. The Fujisaki model describes the pitch contour of a speech signal through the parameters base frequency, phrase commands, and accent commands. While the base frequency represents the minimum F0 value in the signal, the phrase commands describe the slowly varying components and the accent commands indicate local peaks in the contour.

The Fujisaki parameters were assessed for four independent auditory evaluated databases consisting of synthetic speech generated by over 20 different text-to-speech (TTS) systems. The prosody generation techniques of these systems is unknown to us, i.e. it may happen that the systems base their prosody on a Fujisaki-like model or not. The extracted parameters were used to calculate 47 features (e.g. mean distance between phrase commands, variance of accent command amplitudes, etc.).

A stepwise multiple linear regression of these features with the overall quality judgement (MOS) as the response variable led to one quality prediction model per gender. A leave-one-out cross-validation showed the stability of both models. The Pearson Correlation R between predicted MOS and auditory MOS was computed per gender and database. The mean correlation reached a value of $R > .50$.

Even though, the computed Fujisaki features do not fully capture the auditory quality of TTS stimuli both models will be helpful for predicting TTS quality. Especially, in combination with other features an increase in accuracy is to be expected.

1 Introduction

The constant improvements of TTS systems over the past decade lead to synthetic voices that no longer remind listeners of the robot-like voices from the eighties, but of real human beings. Even though they can still easily be distinguished from human-produced speech the increase of naturalness made it possible to use TTS for every-day applications like email readers, information services, or smart-home assistants. Even more challenging tasks such as audiobooks come into focus. The continuous improvement of TTS systems and their application to new use cases always requires a frequent evaluation of their quality. As a consequence, methods for efficiently assessing overall quality as well as its underlying perceptual dimensions are of great interest.

Depending on which aspect of the system is to be evaluated, different types of listening tests are recommended. Most listening tests however, e.g. the method described in the ITU-T Rec. P.85 [1], are used to quantify the inherent quality dimensions of the synthesized speech signals through ratings on multiple scales, such as naturalness, pronunciation, intonation, etc. The major drawback is that such listening tests are very cost-intensive as well as time-consuming which makes it hard for developers of synthetic speech to evaluate the quality of their systems after each step in the development process. Therefore, instrumental methods that predict the quality of synthetic speech without the need for actual human listeners could accelerate this process. Such methods have been implemented and tested on different TTS databases. In [2] and [3] three different approaches were used to predict TTS quality: a HMM-based feature comparison, a linear predictor comprising 45 internal features of the ITU-T Rec. P.563 [4], and one linear predictor based on a set of about 1500 general speech features. Each of these approaches achieved strong correlations with the auditory MOS for some of the TTS databases. Nevertheless, all of them scored weaker on at least one database or for one gender. In [5] formal parameters of speech prosody were analyzed towards their usefulness to estimate perceptual quality of synthetic speech. They describe the F0-contour as well as the temporal structure of speech through statistical features (range, standard deviation, mean, etc.) respectively measures of durational proportionality and variability of vocalic and inter-vocalic intervals. 18 acoustic features were used in a regression analysis. The resulting correlations with auditory ratings were as high as .87.

With these pleasant results other prosody related features are brought into focus. In this paper we present research on the use of the Fujisaki model to estimate the quality of TTS signals. The Fujisaki model extracts the pitch curve of speech signals and reduces its complexity to a minimal set of parameters. These parameters describe the pitch contour of a speech signal through a superposition of a minimum F0 value, a slowly varying component and local peaks in the signal. The parameters are used to derive 47 features as an input for a stepwise multiple linear regression analysis to create one predictor for each speaker gender.

In Section 2 we give a short overview on the Fujisaki model as well as a detailed description of the features derived from the extracted parameters. Section 3 presents the four TTS databases that were used for the development of our predictors. A stepwise multiple linear regression analysis on the computed features is carried out in Section 4. Applying the developed models to our databases, we analyze the performance and robustness of the predictions in Section 5. Finally, in Section 6 we summarize the results.

2 Fujisaki model and features

In this section we present a brief overview of the Fujisaki model as well as a detailed description of the developed features based on the extracted Fujisaki parameters.

2.1 Fujisaki model

The F0-contour of speech signals contributes important non-linguistic information like naturalness and the current emotion of the speaker. Generally, such contours are characterized by a decline from onset towards the end of an utterance. During word accent the F0-contour is superposed by local intonation humps.

The Fujisaki model [6] follows this principle by describing a F0-contour as a superposition of phrase (PC) and accent commands (AC) and an underlying basefrequency (BF). The concept of this model can be seen in Figure 1.

PCs consist of several starting points, each of them with a specific amplitude. Thus, they describe a set of impulses. PC amplitudes as well as the onset time for the first PC of a signal can have a negative sign. ACs consist of starting and ending points that describe a set of stepwise functions. The time within one pair of starting and ending points represents an accented block. In comparison to PCs all AC amplitudes and the onset time of AC are always positive. The BF describes the minimum value of the logarithmized F0-contour throughout the signal.

The PCs and ACs are the input for two critically-damped second-order linear systems to these commands (*phrase control mechanism* respectively *accent control mechanism*). The PCs and

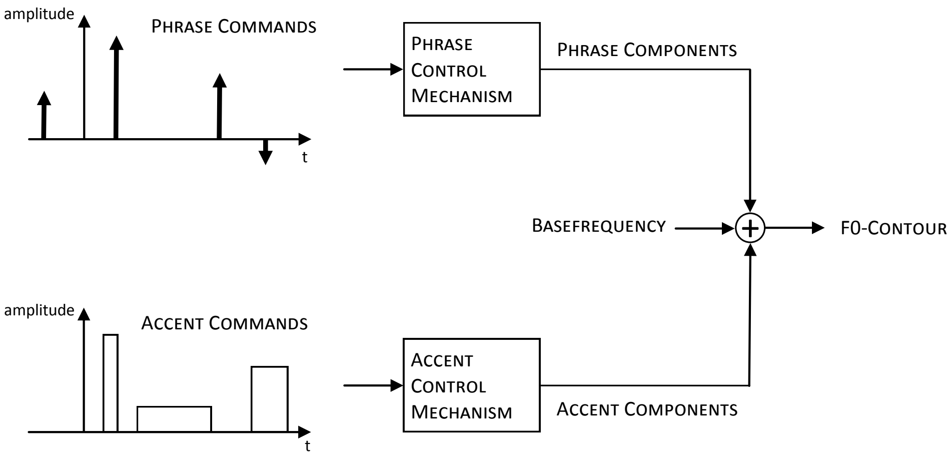

Figure 1: Fujisaki model for the generation of F0 contours.

ACs are assumed to be smoothed by the low-pass characteristics of their respective control mechanisms. The output of those control mechanisms (the *phrase components* and *accent components*) and the BF are then joined to form the pitch curve of an utterance. Thus, this model reduces the complexity of a pitch contour to a minimal set of three parameters (PC, AC, and BF) that still capture the main aspects of the pitch curve.

2.2 Fujisaki features

Keeping in mind that the prosody of synthetic speech is the most relevant aspect when it comes to the impression of naturalness we wanted to use a model that describes the pitch contour as the basis for a quality predictor. Therefore, we used the Fujisaki model implemented especially for the use in German [7] to extract the above-mentioned parameters for all TTS files from the databases mentioned in the following section.

We then computed 47 statistical features based on the extracted Fujisaki parameters. They comprise mean, minimum, maximum values as well as the variance of the extracted parameters. Moreover, we computed several features based on the quantity of increasing/descreasing (in relation to the previous command) PC/AC segments in a signal. All features can be derived from Equations (1) to (10) by combining the terms in curly brackets in every possible way, e.g. one of the features from equation (1) is called *maximum of distances between AC starting points*.

$$\left\{\begin{matrix} mean \\ minimum \\ maximum \\ variance \end{matrix}\right\} of\ distances\ between \left\{\begin{matrix} PC\ starting\ points \\ AC\ starting\ points \\ AC\ ending\ points \\ AC\ starting\ points\ and\ following\ ending\ points \\ AC\ ending\ points\ and\ following\ starting\ points \end{matrix}\right\} \tag{1}$$

$$\left\{\begin{matrix} mean \\ minimum \\ maximum \\ variance \end{matrix}\right\} of \left\{\begin{matrix} PC\ amplitudes \\ AC\ amplitudes \end{matrix}\right\} \tag{2}$$

$$number\ of \left\{\begin{matrix} increasing \\ decreasing \end{matrix}\right\} \left\{\begin{matrix} PCs \\ ACs \end{matrix}\right\} normed\ by\ length\ of\ the\ signal \tag{3}$$

$$relative\ position\ of \begin{Bmatrix} minimum \\ maximum \end{Bmatrix} \begin{Bmatrix} PC\ amplitude \\ AC\ amplitude \end{Bmatrix} \tag{4}$$

$$\begin{Bmatrix} minimum \\ maximum \\ sum\ of\ all \end{Bmatrix} AC\ block(s) \tag{5}$$

$$sum\ of\ all\ AC\ blocks\ normed\ by \begin{Bmatrix} maximum\ amplitude \\ maximum\ AC\ block \end{Bmatrix} \tag{6}$$

$$ratio\ of\ increasing\ and\ decreasing \begin{Bmatrix} PCs \\ ACs \end{Bmatrix} \tag{7}$$

$$quantity\ of \begin{Bmatrix} PCs \\ ACs \end{Bmatrix} normed\ by\ length\ of\ the\ signal \tag{8}$$

$$basefrequency \tag{9}$$

$$sum\ of\ PC\ amplitudes\ normed\ by\ maximum\ amplitude \tag{10}$$

3 Databases

3.1 Test 1

Test 1 has been carried out at the Institute of Phonetics and Digital Speech Processing at Christian-Albrechts-University of Kiel, Germany (see [8] for details). The stimuli were generated by 3 commercial and 3 academic TTS systems. Female and male voices of each system were used to synthesize 5 stimuli per gender and system. The synthesized speech signals consist of 2 utterances separated by a silence interval of about 2s. The stimuli have an average duration of 12s. Additionally, 2 natural speakers per gender were integrated into the listening test. All speech signals were bandpass-filtered (300-3400Hz) and level normalized to an active speech level of -26dBov prior to listener presentation. 17 test participants (10 female, 7 male) were acquired; all of them were German students and their age ranged from 20-26. Each of them rated all 80 stimuli on 8 scales, most of them were recommended by the ITU-T Rec. P.85 [1].

3.2 Test 2

The second test [9] was carried out at the Quality and Usability Lab, TU Berlin. 6 different TTS systems with female and male voices were used to synthesize 5 samples per system. Additionally, natural speech files with female and male voices were included. The average duration was 7-8s. All stimuli were pre-processed as in Test 1, coded-decoded with log PCM according to ITU-T Rec. G.711, and presented to 25 test participants (12 female, 13 male, mean age: 25.8 years) in a soundproof booth. The listening test closely followed the procedure described in ITU-T Rec. P.85, using 4 rating scales to measure the quality, naturalness, etc. of the presented signals.

3.3 Test 3

The Test 3 database resulted from a study [10] in which the inherent quality dimensions of state-of-the-art TTS systems were investigated. 14 female and 15 male synthesizers were used to generate 2 samples per system. All stimuli were downsampled to fs=16kHz and level normalized to -26dBov prior to listener presentation. The average duration was 9-10s. 30 listeners (15, female, 15 male, mean age: 27.9 years) rated all signals on 16 attribute scales that were developed during two extensive pretests and on an overall quality scale.

3.4 Test 4

The forth database was gathered during further research on quality dimensions of synthetic speech. Therefore 30 female and 27 male stimuli all synthesized by different TTS systems were rated on the same scales as described in Test 3. Prior to listening the stimuli were downsampled to 16kHz and level normalized to -26dBov. 12 test participants (5 female, 7 male, mean age: 27 years), 5 expert listeners from Telekom Innovation Laboratories and 7 naïve subjects took part in the test. All of them were native German speakers. The stimuli had an average duration of 5s and were presented via head-phones (Sennheiser HD 485) and a high-quality sound device (Roland Edirol UA-25) in a quiet listening environment.

3.5 Differences between databases

Even though, all databases were generated via similar test procedures, there are differences that have to be taken into account during the following steps and the interpretation of the results.

- Test 1 and 2 used Absolute Category Rating (ACR) and the MOS overall quality scale while Test 3 and 4 used continuous scales with a range of 1 to 7.
- Test 1 and 2 used natural speakers as reference stimuli. This leads to a compression of the range in which TTS stimuli are rated.
- All databases consist of partially different TTS systems. Hence, the ratings in one database always also depend on the range of quality of TTS systems in it.
- The mean duration of stimuli varies between databases from 5s to 12s.

4 Prediction models

We aim to develop one overall quality predictor based on the presented Fujisaki features that best estimates the auditory MOS of all four databases described in the previous section. Therefore, those databases had to be merged. Due to the different rating scales we decided to scale the ratings of Test 3 and 4 to the standard MOS scale range (1 to 5). Moreover, we omitted the natural speech files from Test 1 and 2 because we are only interested in the evaluation of synthetic speech.

As we learned from previous research [2] the prediction efficiency of most features varies highly between genders. Hence, we conducted one stepwise multiple linear regression analysis for each gender. The auditory MOS of all four databases were used as response variable while the 47 Fujisaki features described in Section 2 were used as predictors.

For both genders one significant model could be created. In Table 1 we list the selected features for the female model, its beta values (B), their standard errors (SE B), and their standardized values (β). The four features denote the number of decaying ACs normed by the length of the speech signal (*num dec ac norm length*), the basefrequency of the signal (*basefrequency*), the minimum distance between ending points and the following starting point of the ACs in a signal (*min dist ac ep sp*), and the mean distance between PC starting points (*mean dist pc sp*). Even though the root mean square error (RMSE) for the female predictor is fairly low ($RMSE_f = 0.52$) the model only accounts for 26% of the variablity in the outcome.

Table 1: Results of stepwise multiple linear regression analysis for female voices. $R^2 = .26$.

FEATURE	B	SE B	β
constant	2.073	0.539	
num dec ac norm length	0.902	0.248	$.304^{***}$
basefrequency	-0.008	0.003	$-.212^{*}$
min dist ac ep sp	6.131	2.099	$.248^{**}$
mean dist pc sp	0.300	0.107	$.234^{**}$

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.
Note: see text for explanation of the features.

The male model (Table 2) consists of 5 predictors. These features denote the mean distance between PC starting points (*mean dist pc sp*), the quantity of ACs normed by the length of the speech signal (*quantity ac norm length*), the mean amplitude in the ACs in a signal (*mean ac amp*), the maximum distance between the ending point and the following starting point of the ACs in a signal (*max dist ac ep sp*), and the sum of all AC blocks in a signal (*sum ac blocks*). The RMSE for this model is on the same level as the RMSE for the female predictor ($RMSE_m = 0.48$) however, the male model accounts for 39% of the variablity in the outcome. Taking a look at both models reveals that the feature *mean dist pc sp* is the only item that shows up in the female as well as the male predictor.

Table 2: Results of stepwise multiple linear regression analysis for male voices. $R^2 = .39$.

FEATURE	B	SE B	β
constant	1.135	0.407	
mean dist pc sp	-0.354	0.089	$-.334^{***}$
quantity ac norm length	0.790	0.171	$.404^{***}$
mean ac amp	4.380	1.030	$.512^{***}$
max dist ac ep sp	0.274	0.072	$.428^{***}$
sum ac blocks	-0.519	0.208	$-.302^{*}$

$^{*}p < .05$. $^{***}p < .001$.
Note: see text for explanation of the features.

To test for over-fitting effects a leave-one-out cross-validation was conducted. The R^2 values for both models could be confirmed. The RMSE showed a minor increase for both the female ($RMSE = 0.55$) and the male ($RMSE = 0.51$) predictor. Thus, both models can be accounted to be stable.

5 Results and discussion

We used both models to compute predicted MOS for all available TTS files. As a measure of accuracy we report on the Pearson correlation coefficient R between predicted MOS and auditory MOS per database and gender and the RMSE. The achieved correlations can be seen in Table 3.

Table 3: Pearson Correlation between predicted MOS and auditory MOS for each database.

	FEMALE		MALE	
DATABASE	R	RMSE	R	RMSE
Test 1	$.48^{**}$	0.44	$.58^{**}$	0.48
Test 2	.05	0.61	$.48^{*}$	0.55
Test 3	$.66^{**}$	0.65	$.60^{**}$	0.62
Test 4	$.61^{**}$	0.79	$.75^{**}$	0.70

$^{*}p < .05$. $^{**}p < .01$.

The results for the female stimuli show a strong correlation for the two more complex databases (Test 3 and 4) and a medium correlation for Test 1. For the female speech files from Test 2 no significant correlation could be achieved. The results for the male files are superior to those of the female databases: for Test 1, 3, and 4 strong correlations could be achieved while Test 2 still reaches $R = .48$.

When comparing the results across databases the correlations for Test 3 and 4 stand out with $R \geq .60$. The lowest correlations for female and male data were achieved on the Test 2 database. For most databases both predictors achieved strong correlations even though the four databases differ from each other in many ways (see Section 3.5). When comparing the results for Test 1 and 2 with the results for Test 3 and 4 it is striking that these correlations are on a lower level. The cause could be that the first two databases contain natural speech stimuli while the others do not. This circumstance might well have led to a relatively poorer rating of the TTS stimuli in those databases during the listening tests than if there would not have been any natural references in them. The same accounts for the various different TTS systems in all four databases. The combinations in each of them also influence the ratings of the stimuli. Moreover, the average length of the TTS signals differs strongly between databases. Until now, we have not determined a minimal required signal length for a reasonable prediction accuracy. But we observed that previous prediction models showed problems with shorter signals of about 3-4s.

Furthermore, the R^2 values for both models do not account for more than 40% of the variablity in the outcome. Thus, to reliably estimate the quality of TTS systems additional features will be necessary.

Keeping in mind that the pitch curve described by the Fujisaki parameters is a decisive factor for the impression of naturalness of synthetic speech we used the presented method to create two predictors with the auditory naturalness ratings as target variable. Surprisingly, the achieved correlations lag far behind the values from the MOS predictor. Therefore, we concentrated on the results from the MOS prediction models.

6 Conclusion

We used the Fujisaki model implemented by Mixdorff [7] to extract parameters for four German TTS databases with over 200 samples. From these parameters 47 statistical features were derived. We conducted one stepwise multiple linear regression analysis with these features as predictors and the auditory MOS as response variable for female and male data. Two stable models could be constructed depending on four features for the female data and five features for the male data. Both models have proven stable with only minor changes in R^2 in a leave-one-out cross-validation.

We computed Pearsons correlation coefficient between the predicted MOS and the auditory

MOS for each database and gender. With the exception of the female data from Test 2 we reached correlations between .48 and .75.

Though the developed models are not able to predict the quality of the TTS signals from all databases with a sufficient accuracy we are confident that the results will contribute to future quality prediction approaches. Especially, in combination with other features we expect a further increase in predictive power.

7 Acknowledgements

The present study was carried out at Deutsche Telekom Laboratories, Berlin. It was supported by the Deutsche Forschungsgemeinschaft (DFG), grants MO 1038/11-1 and HE 4465/4-1.

References

[1] ITU-T Rec. P.85: *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices.* Geneva: International Telecommunication Union, 1994

[2] Möller, S. ; Hinterleitner, F. ; Falk, T.H. ; Polzehl, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems. In: *Proceedings of the 11th Annual Conference of the ISCA (Interspeech 2010). International Speech Communication Association (ISCA)* (2010), S. 1325–1328

[3] Hinterleitner, F. ; Möller, S. ; Falk, T.H. ; Polzehl, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009. In: *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)* (2010)

[4] ITU-T Rec. P.563: *Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony.* Geneva: International Telecommunication Union, 2004

[5] Norrenbrock, C. ; Hinterleitner, F. ; Heute, U. ; Möller, S: Instrumental Assessment of Prosodic Quality for Text-To-Speech Signals. In: *IEEE Signal Processing Letters* 19 (2012), S. 255–258

[6] Fujisaki, H.: Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. Acoustical Analysis and Physiological Interpretations. In: *STL-QPSR* Vol. 22 (1981), S. 1–20

[7] Mixdorff, H.: *MANUAL for the FujiParaEditor - An Interactive Tool for Extracting Fujisaki Model Parameters.* http://public.beuth-hochschule.de/ mixdorff/thesis/fujisaki.html, 2010

[8] Seget, K.: *Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren (Study of Perceptual Quality of Text-to-Speech Systems).* Diplomarbeit, Lehrstuhl für Netzwerk- und Systemtheorie, Christian-Albrechts-Universität Kiel, 2007

[9] Hinterleitner, F.: *Vorhersage der Qualität synthetischer Sprache mittels eines signalbasierten Maßes.* Magisterarbeit, Quality and Usability Lab, TU Berlin, 2010

[10] Hinterleitner, F. ; Möller, S. ; Norrenbrock, C. ; Heute, U.: Perceptual Quality Dimensions of Text-to-Speech Systems. In: *Proc. of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)* (2011), S. 2177–2180