Quality Analysis of Macroprosodic $F_0$ Dynamics in Text-to-Speech Signals

Christoph R. Norrenbrock$^1$, Florian Hinterleitner$^2$, Ulrich Heute$^1$, Sebastian Möller$^2$

1Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel, Germany
2Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
{cno, uh}@tf.uni-kiel.de, {florian.hinterleitner, sebastian.moeller}@telekom.de

Abstract

We present a study on the relation between fundamental frequency ($F_0$) and its perceptual effect in the context of text-to-speech (TTS) synthesis. Features that essentially capture the intonational (macro-prosodic) properties of spoken speech are introduced and analysed with regard to the following questions: (i) How does the prosodic variation of TTS signals differ from natural speech? (ii) Is there a functional relationship between the prosodic variation of TTS signals and its perceived quality? In answering these questions we present novel approaches for the construction of non-intrusive quality estimators. The results reveal a substantial degree of systematic influence of prosodic variation on TTS quality.

Index Terms: Speech quality, instrumental quality assessment, text-to-speech (TTS), prosody.

1. Introduction

The wish to evaluate speech quality with the aid of instrumental measures, that is to provide an estimate of subjective speech quality by algorithmic means, concerns most researchers who modify a speech signal, for example by noise reduction or coded transmission. A special case of this objective applies to text-to-speech (TTS) research, where a generated speech signal needs to be evaluated with respect to its perceptual quality. In the literature, efforts to develop a reliable instrumental measure of TTS signal quality are mostly directed towards the detection [1] and perceptual quantification [2] of concatenation artefacts, since most state-of-the-art TTS techniques are based on concatenation of speech units with the inherent risk of spectral mismatch at their boundaries. In turn, numerical measures, capable of estimating the perceptual effect of a given unit joint, could lead to improved smoothing techniques for concatenated synthesis [3]. Naturally, this research line is mostly based on the analysis of single spoken words or rather short stimuli. However, in case of longer sentences (e.g., 10 s), the influence of prosody, that is the joint variation of $F_0$, duration, and intensity in spoken speech [4], becomes highly relevant. Multiple studies give rise to the assumption that, at least for the German language, the perceived quality (i.e., naturalness) of TTS signals is mainly determined by their prosodic quality [5, 6]. Apparently, a main challenge is to set (or modify) $F_0$ in a way that reliably satisfies linguistic needs of the text while ensuring perceptual naturalness at the same time. In this context, furthering the understanding of prosodic adequacy is particularly important to improve TTS systems in terms of their overall perceptual quality. The establishment of instrumental evaluation techniques for prosodic quality however, appears to be a challenging task since there is no unique solution to set prosodic parameters to some correct values, or, more generally speaking, prosody prediction from text is a heavily under-specified problem [7]. One rare effort to tackle this problem is documented in [8]. In this study, an intrusive (reference-based) approach was adopted for evaluating the perceptual difference of aligned $F_0$ contours by means of their numerical difference (synthetic vs. natural). The results suggest that differences of $F_0$ curves are not perceived in a linear way, as gross errors in intonation can overshadow smaller differences in localised $F_0$ [8]. Taking these experiences into account, we propose to drive towards non-intrusive evaluation of $F_0$. First, statistical differences between synthetic and natural speech signals are explored with the aim of identifying systematic deficits of prosodic $F_0$ variation that can be extracted from the acoustical signal alone. Second, insights into the construction of a non-intrusive quality measure are given that tie up to a recent study headed towards instrumental measurement of prosodic TTS quality [9].

The paper is organised as follows: In Section 2, a set of prosodic parameters, also denoted as features, is introduced. The used speech material is described in Section 3. In Section 4, we compare parameter statistics for synthetic and natural speech signals. In Section 5, an approach to perceptual modelling is presented. The results are discussed in Section 6. The paper closes with conclusions.

2. Prosodic Parameters

The set of prosodic features is denoted as $X = \{x_1, \ldots, x_i, \ldots, x_j\}$. Individual features $x_i = \{x_{i1}, \ldots, x_{in}, \ldots, x_{iN}\}$ all manifest as one time-aggregated scalar $x_{in}$ per signal (stimulus). $N$ is the number of observations (stimuli) of the dataset under test (see Section 3). The choice of features ($I = 14$) was guided so as to allow for general insight into $F_0$ variation in TTS on one hand, and to focus on those features which have shown best performance for perceptual quality modelling on the other hand, see [9].

Let $F_0(l, v)$ be the pitch contour of the $l$-th voiced segment, characterised by $F_0(l, v) \neq 0$, with $l = 1, 2, \ldots, L$ and $v = 1, 2, \ldots, V_l$. $L$ denotes the number of voiced segments per signal and $V_l$ is the number of $F_0$ samples extracted at a rate of 100 Hz using Praat [10]. Median filtering is applied to alleviate outliers. The following well-known parameters are considered first: $\Delta F_0$ (range), $\sigma F_0$ (standard deviation), and $\overline{F}_0$ (mean). These parameters are applied signal-wise to the voiced-only $F_0$ curve (i.e., $L = 1$) which is gauged to the logarithmic semitone (st) scale [4], referenced to the minimum $F_0$ value:

$$F_0(l, v) \text{ [st]} = 12 \cdot \log_2 \frac{F_0(l, v) \text{ [Hz]}}{\min_{l,v} F_0(l, v) \text{ [Hz]}}.$$  

(1)

This normalisation can be interpreted as an $F_0$-specific speaker normalisation, with the additional effect that the range is equivalent to the maximum $F_0$. 

10.21377/Interspeech.2012-156
INTERSPEECH 2012

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$\text{MALE}$</th>
<th>$\text{FEMALE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR, DR</td>
<td>0.325 [µ/10 ms]</td>
<td>0.55 [µ/10 ms]</td>
</tr>
<tr>
<td>VR</td>
<td>0.65 [µ/10 ms]</td>
<td>1.1 [µ/10 ms]</td>
</tr>
</tbody>
</table>

Table 1: Settings of threshold $\xi$.

From the literature [4], it is known that the perception of $F_0$ (e.g., pitch distance and pitch change) is substantially influenced by perceptual auditory thresholds. Moreover, intonational properties of $F_0$ curves can be described by piecewise linear functions (stylisation). Following this line of research, we propose nonlinear $F_0$ parameters, calculated from the slope $m_{reg}$ of the least-squares regression line fitted through the segments $F_0(l, v)$. The *peakedness ratio* (PR) is defined as the relative number of segments per signal whose magnitude of $m_{reg}$ is above a threshold $\xi$.

$$\text{PR} = \frac{1}{L} \sum_{l=1}^{L} \delta_\xi \left( m_{\text{reg}} \{F_0(l, v)\}_{v=1,\ldots,v_L} \right) ,$$

with the step function $\delta_\xi(x)$ defined as:

$$\delta_\xi(x) = \begin{cases} 1, & \text{for } x > \xi \in \mathbb{R}^+ \\ 0, & \text{else} \end{cases} .$$

(2)

The *drop ratio* (DR) expresses the fraction of declining segments:

$$\text{DR} = \frac{1}{L} \sum_{l=1}^{L} \delta_\xi \left( -m_{\text{reg}} \{F_0(l, v)\}_{v=1,\ldots,v_L} \right) .$$

(3)

Finally, we take into account the *variability ratio* (VR), which gives the relative number of segments with a mean derivative above $\xi$:

$$\text{VR} = \frac{1}{L} \sum_{l=1}^{L} \delta_\xi \left( \frac{1}{v_L-1} \sum_{v=1}^{v_L-1} |F_0(l, v) - F_0(l, v + 1)| \right) .$$

(4)

The thresholds are set in accordance with [9], gender-discriminative to the values given in Table 1. (1) is not used for features PR, DR, and VR as their ability to describe the perceptual effect has not been found to significantly improve through logarithmic compression. In the scope of this study, we further include the per-signal means and standard deviations of the arguments of the step function in (2), (4), and (5), which are beyond the corresponding thresholds. These arguments are the regression slopes $m_{\text{reg},F}$, $m_{\text{reg},D}$ and the mean derivative $|F_{0,v}'|$ per voiced segment, respectively. In [9] eleven rhythm (timing) parameters have been investigated concerning their capabilities to describe TTS quality. To reduce overload, we consider 2 parameters here that have proven robust and are purely derived from $F_0$: The percentage of total voiced duration per signal (%V) and the average duration of voiced segments $T_{\text{voi}}$ in milliseconds.

### 3. Speech Databases

Details of two speech signal databases are given in the following. First, a broad TTS database for which extensive auditory tests have been conducted is described. Second, a natural speech corpus is presented that is used for reference purposes.

### 3.1. TTS Database

The perceptual quality of TTS signals has been analysed within the framework of a semantic differential, carried out at T-Labs Berlin, Germany [5]. Ten German sentences were prepared such that none of them contained words out of the scope of German dictionaries (e.g., proper names). The average spoken stimulus length was about 10 s. For synthesis, 14 and 15 different state-of-the-art TTS systems (concatenation and HMM-based) with female and male voices, respectively, were used. All speech files were downsampled to 16 kHz and level-normalised to -26 dBov. Thirty naïve listeners rated $N = 30$ stimuli per gender on 16 attribute scales. These scales have been determined in two separate pretests [5]. Each TTS system was represented by at least 2 randomly chosen stimuli. Factor analysis revealed a three-dimensional quality model. The dimensions (factors) are named (1) naturalness, (2) disturbances, and (3) temporal distortions. In the scope of this study we consider the pre-stimulus factor scores from a Principal Component Analysis (Varimax rotation) as target variable $y_f = \{y_{1f}, \ldots, y_{5f}\}$, $f = 1, 2, 3$. The factor ordering (1-3) thus corresponds to their perceptual importance in terms of explained variance of the auditory ratings.

### 3.2. Natural Speech Database

In order to allow for a comparison between synthetic and natural speech, the 10 sentences that have been synthesised (see Section 3.1), have also been recorded from 4 male and 4 female human (amateur) speakers. This yields 40 signals per gender. All files are processed to have a sampling frequency of 16 kHz and a normalized speech level of -26 dBov.

### 4. Comparison of TTS vs. Natural Speech

The feature set from Section 2 is analysed in a comparison between synthetic and natural speech using classical statistical figures. In Table 2, arithmetic means and standard deviations of features $x_i$ are denoted as $\mu$ and $\sigma$, respectively, and given separately for both genders. The complete speech material described in Section 3 has been used. The statistical figures reveal a fairly clear message: $F_0$ dynamics in TTS are well below those of natural speech. Nearly all parameters show a remarkable difference in the mean, qualitatively consistent for both genders.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\mu$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta F_0$</td>
<td>14.9</td>
<td>8.3</td>
<td>19.2</td>
<td>6.6</td>
</tr>
<tr>
<td>$\sigma F_0$</td>
<td>2.5</td>
<td>0.8</td>
<td>3.5</td>
<td>0.6</td>
</tr>
<tr>
<td>$F_0$</td>
<td>5.3</td>
<td>1.6</td>
<td>9.0</td>
<td>2.3</td>
</tr>
<tr>
<td>$</td>
<td>m_{\text{reg},F}</td>
<td>$</td>
<td>1.5</td>
<td>0.8</td>
</tr>
<tr>
<td>$</td>
<td>m_{\text{reg},D}</td>
<td>$</td>
<td>1.5</td>
<td>2.2</td>
</tr>
<tr>
<td>$\text{PR}$</td>
<td>0.6</td>
<td>0.2</td>
<td>0.8</td>
<td>0.7</td>
</tr>
<tr>
<td>$\text{DR}$</td>
<td>-1.2</td>
<td>-2.3</td>
<td>-2.3</td>
<td>-2.3</td>
</tr>
<tr>
<td>$\sigma</td>
<td>m_{\text{reg},F}</td>
<td>$</td>
<td>1.6</td>
<td>3.2</td>
</tr>
<tr>
<td>$\sigma</td>
<td>m_{\text{reg},D}</td>
<td>$</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>$</td>
<td>F_{0,v}'</td>
<td>$</td>
<td>1.5</td>
<td>0.9</td>
</tr>
<tr>
<td>$\text{VR}$</td>
<td>1.6</td>
<td>2.0</td>
<td>2.0</td>
<td>1.2</td>
</tr>
<tr>
<td>$T_{\text{voi}}$</td>
<td>0.7</td>
<td>0.3</td>
<td>0.9</td>
<td>0.1</td>
</tr>
<tr>
<td>$% \text{V}$</td>
<td>16.5</td>
<td>5.0</td>
<td>17.0</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 2: Arithmetic means $\mu$ and standard deviations $\sigma$ of features $x_i$, given per gender for TTS signals and natural speech signals.
Regarding the standard deviations, we also note increased values for male TTS compared to natural speech. In contrast, the female standard deviations are often reduced. In comparison to natural speech, this indicates that \( F_0 \) modelling for female voices shares somewhat greater similarity across TTS systems than for male voices. Concerning the timing parameters \( T_{\text{vol}} \) and \%V, values for male and female signals are very similar. Increased \( \sigma \) values for TTS indicate a broader range compared to natural speech.

### 5. Perceptual Modelling

The central question of this study is whether appropriate perceptual models could be constructed from the prosodic parameters introduced in Section 2. In answering this question we first analyse the relevance of core parameters within the quality space of the TTS database. Second, we apply sequential feature selection in combination with regression models to investigate the joint relationship between features and perceptual quality. Furthermore, cross-validation is adopted to give a realistic estimate of prediction accuracy rather than regressive fit. Finally, two experimental research lines are presented: These are (i) a reference mapping and (ii) a feature extension using second-order basis expansions. Pearson’s correlation coefficient \( R \) is used as a measure of linear dependency between features, perceptual ratings, and predictions thereof.

#### 5.1. Analysis of Quality Space

Table 3 gives the correlation coefficients between factor scores \( y_f \) and selected features \( x_i \), separately for each gender. Bold values are significant \((p < 0.05)\).

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( y_4 )</th>
<th>( y_5 )</th>
<th>( y_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta F_0 )</td>
<td>0.58</td>
<td>0.34</td>
<td>0.37</td>
<td>0.55</td>
<td>-0.03</td>
<td>0.28</td>
</tr>
<tr>
<td>( \sigma F_0 )</td>
<td>0.54</td>
<td>0.19</td>
<td>0.34</td>
<td>0.39</td>
<td>-0.02</td>
<td>0.14</td>
</tr>
<tr>
<td>( T_{\text{vol}} )</td>
<td>0.45</td>
<td>0.18</td>
<td>0.26</td>
<td>0.61</td>
<td>-0.15</td>
<td>0.45</td>
</tr>
<tr>
<td>PR</td>
<td>0.76</td>
<td>0.21</td>
<td>0.49</td>
<td>0.52</td>
<td>0.14</td>
<td>-0.04</td>
</tr>
<tr>
<td>DR</td>
<td>0.71</td>
<td>0.27</td>
<td>0.51</td>
<td>0.58</td>
<td>-0.07</td>
<td>0.11</td>
</tr>
<tr>
<td>VR</td>
<td>0.74</td>
<td>0.15</td>
<td>0.47</td>
<td>0.61</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>( T_{\text{vol}} )</td>
<td>-0.43</td>
<td>0.07</td>
<td>-0.61</td>
<td>-0.43</td>
<td>0.12</td>
<td>-0.60</td>
</tr>
<tr>
<td>%V</td>
<td>-0.56</td>
<td>0.07</td>
<td>-0.52</td>
<td>-0.34</td>
<td>0.27</td>
<td>-0.43</td>
</tr>
</tbody>
</table>

Table 3: Correlations between features and factor scores.

The interpretation is as follows: For both genders most features show a marked relationship with the first factor (naturalness). Regarding the pure \( F_0 \) parameters, the positive correlation \((R > 0)\) most likely reflects the negative perceptual impact of lacking \( F_0 \) dynamics in TTS as documented in Table 2. No significant correlation appears with the second factor. Concerning the third factor, significant cross-correlations are notable especially for DR, PR and VR. However, this applies only for the male stimuli, where a reasonable number of low-quality stimuli with temporal distortions also share intonational deficits. This is not the case for the female stimuli. In turn, timing parameters \( T_{\text{vol}} \) and \%V correlate negatively. The marked correlations with factor 3 (temporal distortions) reflect the effect of polyphony (impression of multiple voices) [5] which can be caused by unnatural variations in speed, lacking speech pauses, and their lengths. Thus, perceptual categories like prosodic punctuation or spoken legato load on factors 1 and 3. In summary, these results suggest the interpretation of naturalness as a foremost prosodic quality dimension. This is also in line with high factor loadings on suprasegmental attributes such as accentuation and rhythm, as reported in [5].

#### 5.2. Reference Mapping

An intuitive way of modelling the naturalness gap between TTS and natural speech is to apply a reference mapping of the features. This provides for the assumption of an “internal” reference of natural speech that the listening test subjects use unconsciously when rating speech quality. In the scope of this study, 2 classical magnitude mappings are considered:

\[
X_{\text{map,lin}} = |X - X_{\text{ref}}|,
X_{\text{map,log}} = \log \left( \frac{X}{X_{\text{ref}}} \right)
\]  

(6)

The reference matrix \( X_{\text{ref}} \) comprises the per-gender means for natural speech (see Table 2). In the following all features are normalised using a modified z-score:

\[
X_{\text{norm}} = \frac{X - X_{\text{ref}}}{\sigma_{X_{\text{ref}}}} + \mu_{\text{norm}}.
\]

(7)

The arbitrary target mean of all features \( \mu_{\text{norm}} \) is set to 0. Matrix divisions in (6) and (7) are realised element-wise.

#### 5.3. Feature Expansion

A controlled shift towards nonlinear feature combination is performed by expansion of the raw feature matrix \( X \) [11]:

\[
X = (h_1(X))|_{x=1}.
\]

(8)

Basis expansion functions \( h_1(X) \) are chosen as:

\[
h_1(X) = X, \quad h_2(X) = x_1 \cdot x_c, \quad h_3(X) = \frac{x_i}{x_c}.
\]

(9)

Function \( h_1 \) simply reproduces the input matrix. Functions \( h_2 \) and \( h_3 \) denote the construction of cross products and cross divisions of all features, respectively.

#### 5.4. Cross-Validated Feature Selection

Repeated sequential feature selection is used in connection with \( K \)-fold cross-validation [11, 9] as follows: For each trial \( m \), a random partitioning of the \( N \) observations into \( K \) disjoint subsets is performed. Forward selection is conducted using the \( K \) training sets \( T_r^{(k)} \) of size \( N \langle 1 - 1/K \rangle \). The \( k \)-th test sets of size \( N/K \) are denoted as \( T_t^{(k)} \). For each temporary feature set \( F_m \subset X \), a linear least squares regression model, acting as a composite quality estimator, is evaluated:

\[
y = \hat{f}_m(X) = \beta_0 + \sum_{j=1}^{\tilde{J}} \beta_j x_j, \quad x_j \in F_m.
\]

(10)

This model delivers estimated auditory ratings \( \hat{y} \) from \( X \) by using a subset of \( J \) predictors \( x_j \), weighted by scalar coefficients \( \beta_j \). The correlation \( R_m \) (regressive fit) serves as selection criterion to be maximized across the \( K \) partitions,

\[
R_m(y, \hat{f}_m) = \frac{1}{K} \sum_{k=1}^{K} R \left( y^{(k)}, \hat{f}_m(X^{(k)}) \right),
\]

(11)

where \( y^{(k)} \) and \( X^{(k)} \) denote the training sets according to \( T_t^{(k)} \). Features are added to form improved temporary models \( \hat{f}_m \) as.
long as $R_m$ increases by at least 0.03. The cross-validated correlation, equivalent to prediction performance, is calculated using the trained (fixed) models $f_m$:

$$R_m^{(CV)}(y, \hat{f}_m) = \frac{1}{K} \sum_{k=1}^{K} R(y^{(-k)}, \hat{f}_m(X^{(-k)})).$$ (12)

The $k$-th test sets according to $y^{(-k)}$ are denoted as $y^{(-k)}$ and $X^{(-k)}$. The normalised root-mean-square error $\epsilon_m$ is given by:

$$\epsilon_m(y, \hat{f}_m) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \frac{|y^{(-k)}_i - \hat{f}_m(X^{(-k)}_i)|}{\max(y) - \min(y)}.$$ (13)

Functions max$(y)$ and min$(y)$ give the maximum and the minimum element of $y$, respectively. These measures reflect the robustness of the composite model of trial $m$. Overall, $M = 500$ trials are carried out which ensures sufficient convergence for the case of $N = 30$ and $K = 3$, applicable in this study. For the experiments, we take the naturalness scores as target variable, i.e., $y = y_1$. Figures of merit are the mean correlation $R_m^{(CV)}$, the standard deviation $\sigma R_m^{(CV)}$, and the averaged root-mean-square error $\tau_m$.

### 6. Results and Discussion

Table 4 gives a performance overview of the different feature matrices. It is noted that the use of the “raw” feature matrix $X$ already yields cross-validated correlations in the range of 0.8, which is considered as an excellent value taking into account the ambiguity of speech prosody. A slight improvement can be realised by including extra rhythm parameters [9]. Turning to the feature expansion using $h_1$, $h_2$, and $h_3$, a somewhat improved correlation for the female stimuli is noted. However, for the male stimuli greater instability of the predictions (increased $\sigma R_m^{(CV)}$) indicates possible risks of nonlinear extensions. The result shows that linear superposition of the proposed features already does a good job as no striking information gain is exposed through consideration of second order feature interactions. Following an additional manual inspection of all features, we see no indication for systematic nonlinearities which could be modelled (e.g., by polynomials) for the benefit of substantial improvements of prediction accuracy. Comparing the two reference mappings, the logarithmic mapping clearly outperforms the linear mapping. However, no improvement can be recorded compared to the usage of no mapping. We see the main reason for this result in the remarkable prosodic difference between TTS signals and natural speech signals. The feature values of natural speech are rarely reached by most of the TTS signals. Furthermore, the listening test comprised TTS signals only. At this stage, we conclude that the differences between TTS signals could be better modelled without consideration of a natural reference. Rather, it seems appropriate to regularise principally unbounded features, such as the mean $F_0$ slope, to a fixed range (e.g., $PR \in [0,1]$). This effectively helps to compensate for outliers.

### 7. Conclusions

In this paper, doors towards robust instrumental quality evaluation of TTS signals have been opened. In particular, it has been shown that formal prosodic parameters, derived from $F_0$ only, enable the prediction of perceptual naturalness ratings with a correlation of up to 0.83 which explains nearly 70% of the rating variance. The essential conclusions are:

- Prosodic variation systematically influences perceptual naturalness of TTS signals and imposes a major impact on their overall perceptual quality.
- $F_0$ dynamics are substantially lower in TTS than in natural speech.
- Prosodic features can be used to construct robust linear models of perceptual quality.
- $F_0$-specific quality differences between TTS signals could be better modelled without a natural reference mapping.

For future work, we envisage the exploration of robust acoustical parameters, suitable for a numerical description of additional quality factors in TTS, for example the naturalness of voice and articulation. Hence, a full objective account of TTS quality could be implemented.

### 8. Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants HE 4465/4-1 and MO 1038/11-1.

### 9. References


