Aperiodicity Analysis for Quality Estimation of Text-to-Speech Signals

Christoph Norrenbrock¹, Ulrich Heute¹, Florian Hinterleitner², Sebastian Möller²

¹Digital Signal Processing and System Theory, Christian-Albrechts-University of Kiel, Germany
²Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany

{cno, uh}@tf.uni-kiel.de, {florian.hinterleitner, sebastian.moeller}@telekom.de

Abstract

This contribution presents a new approach towards non-intrusive quality assessment of text-to-speech (TTS) signals. Perturbation measures which capture the degree of excitation-specific aperiodicity in voiced speech are investigated concerning their quality implications in synthesized speech. Based on two independent TTS databases for which formal attribute-based listening tests have been conducted, we show that perturbation measures are sensitive to quality aspects of prosody and voice characteristic. Furthermore a dominant dependency on TTS type, namely non-uniform unit-selection and diphone synthesis, is identified. Yet, considerable differences between male and female TTS samples are recognized, emphasizing the need for gender-specific quality assessment.

Index Terms: speech synthesis, quality prediction, voice quality, pathological voices

1. Introduction

The quality of text-to-speech (TTS) systems has reached a level which makes them attractive for many every-day applications in the area of telecommunication services such as email and SMS readers. Hence, the growing customer circle as well as the TTS engineers impose an increased weight on the question how the quality of TTS systems can be assessed with the aid of instrumental quality predictors where expensive costly listening tests may be avoided. In contrast to natural speech signals, where instrumental quality predictors have reached a reasonable degree of performance, the assessment of synthesized signals of TTS systems remains to be a difficult task. This is because research is mainly concentrated on the development of double-ended (intrusive) quality estimators, where a clean reference signal is compared to its degraded version. However, in the past decade new non-intrusive methods, capable of estimating the quality without a given reference signal, have emerged also, since the growing complexity of modern speech transmission systems does not always allow such a reference to be available. Standardized algorithms such as ITU-T Rec. P.563 [1] and ANIQUE+ [2] have been shown to be performing with considerable reliability when compared to reference-based algorithms. Unfortunately, these tools are not capable of estimating the inherent qualities of TTS systems because they are designed towards the estimation of quality degradations imposed by suboptimal transmission [3].

Recently, promising models have been presented which capture the quality of TTS systems by exploiting a statistical reference for natural speech [4]. Furthermore, the potential usefulness of internal quality parameters from ITU-T Rec. P.563 has been investigated by means of a regression tree where it was shown that TTS quality should principally be measurable [5]. However, there is an evident lack of knowledge about how this quality can be measured by means of reliable acoustic markers. From the wide range of apparent artefacts, a specific and highly complex aspect of the perceived naturalness concerns the characteristic of the voice which can often be described in terms of increased roughness and hoarseness. However, the question of what features any voice excitation signal has to exhibit in order to yield the naturalness of a human voice is not solved satisfactorily. This is because manifold voice-quality variations in terms of phonation mode (laryngealized, modal, and breathy) occur in spoken language which can be analyzed from a spectral and/or time domain perspective by means of measuring period-to-period irregularities (see section 2 and [6]). Technically this phenomenon is denoted as perturbation which is usually associated in context of disordered voices. We will use the term aperiodicity synonymously in this paper.

Despite an impressive amount of research on this topic [7], no established consensus has been found yet. One main reason is certainly the diversity of experimental perspectives. For example, the type of sample to be analyzed, namely sustained vowel versus continuous speech, has strong implications on the results of voice quality evaluation, both in a perceptual (subjective ratings) and signal-based context. Furthermore, the aim of synthesizing highly natural speech as addressed in [6] and measuring the severity of dysphonia in pathological voices by means of perturbation measures (see section 2) is divergent to some extent. An obstacle, applicable in any case, lies in the controversy of finding reliable models of vocal aperiodicity in psychoacoustic voice synthesis experiments [8]. Though there is no doubt that natural speech exhibits at least a small aperiodic component in voiced parts, with distinct effects on voice quality, it is yet difficult to quantify this aperiodicity exactly and hence there exist no undoubted rules how perturbation should be synthesized or integrated into a TTS system in order to yield the wished contribution to naturalness [6].

In light of this ambiguity, the behaviour of voice quality parameters in TTS systems is of special interest. In particular, the state-of-the-art concatenation synthesis, where speech units are connected via pitch-synchronous overlap-add techniques (PSOLA), appears to be rewarding for such an investigation because of the following argument: Since the inventory of such a TTS system consists of prerecorded units from human speakers, it is reasonable to assume that they capture a natural degree of perturbation. However, depending on the synthesis process, some change in the original degree of perturbation can be expected. Hence, our approach is to investigate perturbation measures across different TTS systems (with different voices) and analyze their quality relation. This is in distinct contrast to studies where a well-defined synthetic excitation signal with a controlled amount of perturbation or an elaborate analysis and synthesis method is utilized [9, 8]. With regard to TTS-specific artefacts, we raise the question to what extent degradations in
naturalness due to concatenation synthesis of a-priori “perfect” speech units can be compared to the degradation of naturalness in disordered voices. We present an extended study of the results in [10], incorporating a broader experimental basis and an analysis of the potential insights that can be drawn from aperiodicity analysis. A description of six perturbation measures and details about the implementation are given in sections 2 and 4, respectively. Section 3 reports about the TTS databases, followed by the discussion of the results in section 5. The paper closes with a conclusion and an outlook on ongoing research.

2. Perturbation measures for voice quality assessment

Over the last decades, a vast number of signal-based acoustic markers have been proposed with the unifying aim of delivering some non-intrusive estimate of the overall voice quality or certain aspects (e.g. roughness) regarding the type or severity of dysphonia. However, none of these measures crystalized to be reliable under all conditions, which is documented by at least 60 reports published on this matter [7]. Most studies are bounded to the analysis of sustained vowels, where elaborate diagnosis tools for clinical application have been developed [11]. Yet, it is desirable to assess voice quality directly from spoken language since the voice character, linked to different phonetic modes in connected speech, varies over time, and a corresponding analysis in voiced parts may give a more precise image of how the voice is actually perceived [12]. As a consequence, effort has been undertaken to extent perturbation analysis to this scenario. As no information is available to the authors which of these measures might work for quality estimation of TTS speech we decided to choose four traditional perturbation measures as representatives for time-domain approaches and two similar cepstral measures which have been shown to exhibit superior performance when applied to continuous speech [7].

2.1. Perturbation in time domain

In general, perturbation can be defined as the average deviation of a fundamental-period parameter \( u(n) \) (e.g. period length or energy) from its mean [11]. If \( n \) denotes the \( n \)-th period of a voiced speech segment consisting of \( N \) periods, the Perturbation Factor (PF) is defined as

\[
PF = \frac{1}{N-1} \sum_{n=1}^{N-1} |u(n+1) - u(n)| \times 100\%,
\]

and expressed in percent. A similar definition, incorporating local averaging of length 3, is the Relative Average Perturbation (RAP) [9]:

\[
RAP = \frac{1}{2} \sum_{n=1}^{N-2} \left[ \frac{1}{2} \sum_{i=0}^{2} u(n+i) - u(n+1) \right] \times 100\%.
\]

When \( u(n) \) is equal to the fundamental period length \( T_0(n) \), the above measures quantify jitter and are termed PPF and RAP (prefix ‘P’ for pitch). For energy perturbation (shimmer), \( u(n) \) is equal to the signal energy per period and equations 1 and 2 are denoted as EPF and ERAP (prefix ‘E’ for energy).

2.2. Perturbation in cepstral domain

From a model perspective, a speech signal \( s(t) \) can be described by the convolution of a glottal excitation signal \( v(t) \) with a slowly time-varying impulse response \( v(t) \) of the vocal tract. An elegant way of separating these two signals is given by the real cepstrum \( C(q) \) with the variable \( q \) termed “quefrency”:

\[
C(q) = F^{-1}[\log|F[s(t)]|];
\]

here, \( F \) denotes the Fourier-transform. Low-quefrency cepstral coefficients represent \( v(t) \), corresponding to the envelope of the spectrum. The excitation \( v(t) \) typically appears as a dominant peak (first “rahmonic”) in the cepstrum, at a quefrency equal to the fundamental period length \( T_0 \) of \( v(t) \). The height of this peak relative to a normalizing regression line through the cepstrum can be interpreted as a measure of periodic dominance and is denoted as Cepstral Peak Prominence (CPP). Smoothing across time and quefrency yields the CPPs [12].

3. TTS databases

In order to investigate the behaviour of the presented measures, we use data from two auditory tests which have been conducted using different concatenative synthesis systems and stimuli as well as different test participants, at different institutes in Germany. In the following subsections, we give only the essential details, further information can be drawn from [4].

3.1. Test 1

Test 1 was carried out at Christian-Albrechts-University of Kiel, Germany. It used speech material from six off-the-shelf TTS systems: Three commercial ones (AT&T, MBROLA-based Prosers, and Cepstral) and three from German academic institutions (TU Dresden, TU Berlin, and University of Bonn), all with male and female voices. A total of 10 German speech samples at a sampling frequency of 8 kHz have been generated per TTS system, half for male speakers and half for female ones. All samples were bandpass-filtered (300-3400 Hz) and normalized to an active speech level of -26 dBs~ prior to listener presentation. The listening test procedure closely followed ITU-T Rec. P.85 [13] and was performed in a silent listening room. 17 listeners rated the stimuli on eight different quality scales. For the present contribution we limit our analysis to the ratings of perceived naturalness (NAT) as this scale has the closest relationship with perturbation.

3.2. Test 2

The second test was carried out at Quality and Usability Lab, TU Berlin. It consisted of speech material from 6 TTS systems: AT&T, ATIP Prosers, DRESS, Nuance RealSpeak, MARY, and MBROLA. For each synthesizer, 5 German speech samples of 7-8 s length (sampling frequency of 8 kHz) were generated. The stimuli have been preprocessed as in test 1, coded-decoded with log PCM according to ITU-T Rec. G.711 and presented to 25 listeners. Again, the procedure described in ITU-T Rec. P.85 was followed, using 4 rating scales of which only the pleasantness (PLN) scale will be further analyzed here due to the most pronounced link to perturbation and its synonymous meaning for naturalness.

4. Implementation

All measures are evaluated in voiced-speech segments only for which a pitch (F0) estimation is provided every 10 ms using an autocorrelation-based algorithm [14] from the phonetics software PRAAT [15]. The pitch-synchronized sequence \( T_0(n) \) is calculated using a subsequent epoch-marking algorithm from Boersma as described in the PRAAT manual (cross-correlation...
type). To ensure a meaningful analysis basis the minimum segment length was set to 5 periods. Per-segment perturbation values are then averaged across the signal.

CPP and CPPs values are calculated using Hamming-windowed blocks of variable length covering 4 pitch periods in case of male and 8 in case of female signals, where the corresponding period length of each frame is gained from the interpolated pitch curve of the signal. The gender-specific discrimination proved advantageous compared to a fixed blocklength as in [12], whereby a comparable spectral resolution is still ensured. The frameshift is fixed to 10 ms. The smoothing for the CPPs is applied over time (3 blocks) and quefrency (10 bins). Averaging over all voiced frames yields per-signal values. All measures, apart from CPP (see section 5) are then linearly regressed onto the NAT and PLN scale, respectively for tests 1 and 2, separately for both genders. An overview of the system is given in Figure 1.

5. Results and discussion

For analysis purposes we calculate the Pearson ($R$) and the Spearman-rank ($\rho$) correlation coefficient between the subjective ratings and the perturbation measures. All values are given on a per-stimulus and per-synthesizer basis and are summarized in Table 1.

Minus signs are kept to emphasize consistent inverse linear relationships with respect to the periodicity measures (CPP, CPPs) as opposed to the aperiodicity measures. Considering the male stimuli from tests 1 and 2, a similar behaviour of all perturbation measures can be observed. Whereas jitter (PPF and PRAP) shows medium correlations around 0.6, shimmer measures cannot be classified as significant. For the cepstral markers, CPPs performs better than CPP which is in line with [7, 12], though CPP performs much weaker here, hence we drop CPP in favour of CPPs for our investigation. Turning to the female stimuli of test 1, a consistent picture can be observed with markedly increased performance (e.g. $\rho_{\text{perm}} = -0.89$ per-synthesizer), especially the time domain measures, now including shimmer (e.g. $\rho_{\text{shimmer}} = 0.94$ per-synthesizer), are highly correlated with NAT ratings. However, for the female stimuli of test 2, no consistency with the remaining findings can be noticed.

Apart from these stimuli, the most interesting observation is the positive linear relationship between naturalness/pleasantness and perturbation, a clear contrast to the usual case in disordered-voice diagnostics where higher perturbation typically indicates a less natural and healthy voice. To highlight this phenomenon we analyze the data of test 1 which consists of diphone and unit-selection-type synthesis samples only. The latter is rated better in average due to a larger inventory which effectively reduces the necessary adaption of the units during synthesis. The original degree of perturbation (from the inventory speaker) is thus better retained than in the diphone case. This affects the voice characteristic (e.g. roughness and hoarseness) and thus the perceived naturalness. Furthermore, the prosody generation in diphone synthesis (DS) is more idealized, leading to much smoother pitch curves.

In Figure 2, NAT ratings and selected perturbation measures are plotted as z-scores (mean-compensated and normalized by the standard deviation) against the stimuli. For both genders stimuli 1-15 stem from diphone synthesis and stimuli 16-30 from non-uniform unit-selection systems. Vertical grid lines discriminate between TTS systems (5 stimuli per system). Apart from stimuli 6-10 index ranges correspond to the same TTS system for both genders.

Generally speaking, perturbation and naturalness are relatively lower for DS in the case of both male and female stimuli. In addition, sentence dependency shows limited impact compared to system dependency which is also documented by some higher per-synthesizer correlations. Also, speaker dependency appears to be bounded since female stimuli 6-15 stem from the same TTS system using two different voices. This result offers high significance in light of the different sentences and speakers used. Regarding the PLN ratings of test 2, the same system-specific dependency is observed as well, yet the relation to perturbation is somewhat different for the female stimuli. To our opinion, the NAT/PLN ratings are best correlated with perturbation when the prosody generation is not only natural in terms
of variation but also correct in terms of phonetic context. For the female stimuli of test 2 this discrepancy contributes to the found deviation. Above all, perturbation should be considered as one aspect of naturalness.

We subsume that the interpretation of perturbation measures is more complex in case of TTS signals due to abnormal macro prosody. Their traditional deployment in continuous speech implies a reasonably healthy modulation of speech melody whereupon the potential abnormal perturbation component adds up to still yield higher perturbation values than for sane voices. In TTS speech this assumption does not hold and a clear differentiation between voice and prosody quality with traditional perturbation does not seem feasible at this point. But since both quality aspects are linked to NAT/PLN ratings, perturbation can still be interpreted as a capable estimation instrument with respect to the residual naturalness induced from the inventory. Illustrating this point, Table 2 gives the correlation values and the root-mean-square errors (rmse) of the estimated NAT and PLN ratings for tests 1 and 2 respectively after multivariate linear regression of all perturbation measures (except CPP) onto the corresponding scales.

### Table 1: Correlation between perturbation measures (PTB) and ratings of naturalness and pleasantness for tests 1 and 2 respectively, reported “per-stimulus/per-synthesizer”.

<table>
<thead>
<tr>
<th>PTB</th>
<th>Test 1 Male</th>
<th>Test 1 Female</th>
<th>Test 2 Male</th>
<th>Test 2 Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R$</td>
<td>$\rho$</td>
<td>$R$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>PPF</td>
<td>0.69/0.72</td>
<td>0.60/0.71</td>
<td>0.78/0.87</td>
<td>0.84/0.94</td>
</tr>
<tr>
<td>EPF</td>
<td>-0.04/-0.03</td>
<td>0.07/0.09</td>
<td>0.64/0.77</td>
<td>0.58/0.77</td>
</tr>
<tr>
<td>PRAP</td>
<td>0.62/0.65</td>
<td>0.39/0.37</td>
<td>0.75/0.85</td>
<td>0.82/0.94</td>
</tr>
<tr>
<td>ERAP</td>
<td>-0.15/-0.17</td>
<td>0.02/0.14</td>
<td>0.73/0.83</td>
<td>0.73/0.94</td>
</tr>
<tr>
<td>CPP</td>
<td>-0.36/-0.40</td>
<td>-0.24/-0.31</td>
<td>-0.49/-0.54</td>
<td>-0.35/-0.37</td>
</tr>
<tr>
<td>CPPs</td>
<td>-0.64/-0.69</td>
<td>-0.59/-0.77</td>
<td>-0.65/-0.70</td>
<td>-0.72/-0.89</td>
</tr>
</tbody>
</table>

### Table 2: Correlations of estimated and auditory-test ratings after linear regression of selected perturbation measures (see section 5). The root-mean-square errors (rmse) are given in the last row.

<table>
<thead>
<tr>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td></td>
</tr>
<tr>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>0.81</td>
<td>0.68</td>
</tr>
<tr>
<td>rmse</td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td>0.32</td>
</tr>
<tr>
<td>0.27</td>
<td>0.30</td>
</tr>
</tbody>
</table>

### 6. Conclusion and outlook

We have presented a new approach to instrumental quality prediction of TTS signals. Perturbation markers, indicating the amount of aperiodicity in speech have been investigated on two auditory test databases. The results show that the presented measures are significantly linked to the perceived naturalness, approving their strong potential for quality prediction. Yet, differences with respect to gender suggest a narrower relation to quality. For that reason the identification of TTS-specific quality dimensions by means of a semantic differential is underway [16]. On this basis further insight with respect to the reported findings is expected.

### 7. Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grants HE 4463/4-1 and MO 1038/11-1.

### 8. References