On Prosodic Quality of Text-to-Speech Signals

Christoph R. Norrenbrock[1], Florian Hinterleitner[2], and Ulrich Heute[1]

[1] *Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel, Germany, Email: {cno, uh}@tf.uni-kiel.de*

[2] *Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany, Email: Florian.Hinterleitner@telekom.de*

Introduction

Speech prosody (fundamental frequency (F_0), duration, and intensity) can be identified as one of the major factors that determine the overall perceptual quality of text-to-speech (TTS) signals [1, 2]. Thus, research towards the establishment of objective functional relations between a given F_0 contour and its perceptual effect (e.g. naturalness) is highly relevant, yet a challenging task since prosodic variation in spoken speech depends on multiple interacting aspects, bound to language and speaker. In the context of TTS engineering, this challenge is further tightened. While synthesising a speech signal, an F_0 contour needs to be designed in a way that ensures perceptual naturalness but also satisfies the linguistic needs of the spoken text. We will present a pilot study on the question to what extent quality aspects in synthesised speech signals could be captured by formal prosodic parameters only.

Method

Our aim is to identify *systematic* links between the auditory quality impression of a TTS signal and its acoustical representation. This is the major precondition for robust objective quality evaluation. First, we present a set of formal prosodic parameters which have proven useful within a previous study [3]. Second, we describe the TTS database for which an auditory test has been conducted. Finally, we explain the novel construction method for instrumental composite quality estimators and review the results.

Prosodic Parameters

A total of 18 features are considered which all manifest as one time-aggregated scalar per signal (stimulus). The features belong to two categories. These are (i) F_0 parameters which mainly reflect intonational (macroprosodic) properties and (ii) rhythm parameters which are derived from vocalic and intervocalic duration.

Let $F_0(l,v)$ be the pitch contour of the l-th voiced segment, characterized by $F_0(l,v) \neq 0$, with $l = 1, 2, ..., L$ and $v = 1, 2, ..., V_l$. L is the number of voiced segments per signal and V_l denotes the number of F_0 samples which are extracted at a rate of 100 Hz using *Praat* [4]. Median filtering is used to alleviate outliers. The following well-known parameters are considered first: ΔF_0 (range), σF_0 (standard deviation), and $\overline{F}_0$ (mean). These parameters are applied signal-wise to the concatenated voiced-only F_0 contour, where the pitch values are transformed to the logarithmic semitone scale [5], referenced to the minimum F_0 value.

Inspired by the search for perceptual thresholds of pitch change in speech [5], we propose nonlinear F_0 parameters, based on 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 ξ,

$$\mathrm{PR} = \frac{1}{L}\sum_{l=1}^{L} \delta_\xi \left(\left| m_{\mathrm{reg}}\{ F_0(l,v)|_{v=1,...,V_l} \} \right| \right), \qquad (1)$$

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

$$\delta_\xi(x) = \begin{cases} 1, & \text{for } x > \xi \in \mathbb{R}^+ \\ 0, & \text{else.} \end{cases} \qquad (2)$$

Similarly, the *rise ratio* (RR) denotes the fraction of rising segments:

$$\mathrm{RR} = \frac{1}{L}\sum_{l=1}^{L} \delta_\xi \left(m_{\mathrm{reg}}\{ F_0(l,v)|_{v=1,...,V_l} \} \right). \qquad (3)$$

The *drop ratio* (DR) is defined in the same way, with the slope m_{reg} multiplied by -1. Finally, we use the *variability ratio* (VR), which gives the relative number of segments with a mean derivative above ξ:

$$\mathrm{VR} = \frac{1}{L}\sum_{l=1}^{L} \delta_\xi \left(\frac{1}{V_l - 1}\sum_{v=1}^{V_l-1} |F_0(l,v) - F_0(l,v+1)| \right). \qquad (4)$$

The rhythm (timing) parameters are described in detail in [3]. Important timing parameters are the percentage of total voiced duration and the average duration of voiced segments per signal.

TTS Database

Six off-the-shelf TTS systems, including commercial (AT&T, Proser, and Cepstral) and research systems (DRESS, BOSS, MBROLA), all with male and female voices, have been used to synthesise 10 German speech samples per system, half for male and half for female voices. All samples were bandpass-filtered (300-3400 Hz) and normalized to an active speech level of -26 dBov prior to listener presentation. The sampling frequency was 8 kHz. A formal listening test was carried out at Christian-Albrechts-Universität zu Kiel, Germany. The test procedure closely followed ITU-T Rec. P.85 [6] and was performed in a silent listening room. 17 naïve listeners rated $N = 30$ stimuli (each ca. 12 s duration) per gender on 8 quality scales which are: Overall impression, listening effort, comprehensibility, articulation, naturalness, prosody, continuity/fluency, and acceptance. Apart

Table 1: Figures of merit of composite quality estimators, applied on unseen data partitions (3-fold cross-validation).

Quality scale	Male		Female	
	$\overline{R}_m^{(\mathrm{CV})}$	$\overline{\epsilon}_m$	$\overline{R}_m^{(\mathrm{CV})}$	$\overline{\epsilon}_m$
Naturalness	**0.87**	0.15	**0.86**	0.17
Overall quality	**0.75**	0.18	**0.75**	0.20
Listening effort	**0.69**	0.19	**0.62**	0.22
Continuity/Fluency	**0.74**	0.19	**0.55**	0.26
Acceptance	**0.59**	0.24	**0.67**	0.28
Prosody	**0.42**	0.27	**0.78**	0.20
Articulation	**0.57**	0.24	**0.50**	0.22
Comprehensibility	**0.47**	0.29	**0.35**	0.22

from acceptance (binary scale), all attributes were rated on the absolute category rating (ACR)-scale, ranging from 1 (bad) to 5 (excellent).

Composite Quality Estimator

Sequential feature selection is used in combination with (multiple) regression models as described in [3]. This method has the advantage that a composite quality predictor is constructed using only the most relevant parameters which simplifies the interpretation of the model. Furthermore, we propose a strict cross-validation principle, that involves a random partitioning of the signals into K subsets, $K-1$ of them are used for training, 1 for testing the model. Model training comprises forward feature selection where the regressive fit of the regression model serves as selection criterion. This process is repeated M times to compensate for bias of individual partitionings. In this study, $K = 3$ and $M = 500$ applies. Figures of merit are the mean correlation $\overline{R}_m^{(\mathrm{CV})}$ (Pearson's correlation between estimated and true ratings of the test partitions) and the corresponding root-mean-square error $\overline{\epsilon}_m$, averaged over all trials.

Results and Discussion

From Table 1, we see that the proposed parameters are most suitable for predicting the perceived naturalness of the TTS signals under test. This result is consistent for the voices of both genders. Strikingly, overall quality ranks second which reflects the paramount influence of prosodic characteristics on TTS quality. The fact that prosody ratings cannot be predicted consistently ($R_{\mathrm{male}} = 0.42$ and $R_{\mathrm{female}} = 0.78$) can be explained through the greater difficulty (disagreement) naïve subjects have in rating a somewhat more abstract quality aspect like prosody, at least when compared to naturalness. Finally, speech comprehensibility appears to be least influenced by the proposed features. In this context, it is noteworthy that comprehensibility does not seem to constitute a major quality issue of modern TTS systems any more [1]. However, listening effort, which is expected to be related to comprehensibility, still seems to be a relevant evaluation factor. Considering the importance of individual features for the prediction of naturalness, we find the F_0-dynamic parameter DR to have the greatest impact. This phenomenon can be associated with the *declination effect* that is common in connected speech [5]. The role of the remaining features is discussed in detail in [3].

The most remarkable result of this study can be seen in the arising possibility to robustly estimate TTS-signal quality by means of acoustical parameters only. This result was not expected to this degree, since other quality factors (e.g., voice quality, concatenation artefacts) also play an important role [1]. Moreover it is of note that the consideration of linguistic correctness of the synthesis seems unnecessary, at least for a prediction of reasonable accuracy. This characteristic greatly enhances the use of fully automatic evaluation techniques for TTS signals, possibly also for languages other than German. Finally, we like to point out the inherent need for a reasonable signal length as a necessary precondition for the estimation of a suprasegmental quality aspect like prosody. Often, auditory tests are carried out with rather short stimulus lengths (e.g. 1-5 seconds). This practice needs to be avoided in favour of much longer stimulus lengths (e.g. 10-15 seconds). Only then it is possible to get a realistic impression of the usability of a TTS system in everyday applications such as email and SMS readers. In this context, we believe that the decision about the applicability of TTS systems for telecommunication services could be greatly simplified in the future.

Acknowledgement

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

References

[1] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute, "Perceptual quality dimensions of text-to-speech systems," *Proc. Interspeech 2011*, Florence, Italy, pp. 2177–2180, 2011.

[2] V. Kraft and T. Portele, "Quality evaluation of five speech synthesis systems for german," *Acta Acustica*, vol. 3, pp. 351–366, 1995.

[3] C. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Instrumental Assessment of Prososdic Quality for Text-to-Speech Signals," *IEEE Signal Processing Letters*, vol. 19, no. 5, pp. 255-258, 2012.

[4] P. Boersma and D. Weenik. (2005) Praat, software for speech analysis and synthesis. [Online]. Available: `http://www.praat.org`

[5] J. 't Hart, R. Collier, and A. Cohen, *A perceptual study of intonation.* Cambridge University Press, 1990.

[6] ITU-T Rec. P.85, "A method for subjective performance assessment of the quality of speech voice output devices," *Int. Telecomm. Union, Geneva, 1994.*