Quality Estimation of Text-To-Speech Signals

Christoph Norrenbrock[1], Ulrich Heute[1], Florian Hinterleitner[2], and Sebastian Möller[2]
[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, Deutsche Telekom Laboratories, TU Berlin, Germany*

Introduction

This contribution addresses the search for acoustic quality correlates in synthetic speech signals. The approach consists of determining the degree of aperiodicity during voiced speech, considering the fact that natural speech exhibits at least a small aperiodic component. This lack in periodicity (perturbation) results in a specific shaping of the Fourier spectrum with strong links to perceptual categories such as breathiness, creak, and roughness [1]. In Text-To-Speech (TTS) systems, these noise-like modulations of the fundamental frequency (jitter) and the amplitude (shimmer) are often neglected, thus contributing to the overall unnaturalness of the synthesized signal. Though concatenation synthesis utilizes inventories recorded from human speakers, the natural degree of perturbation captured in the units is altered, depending on unit length and the extent of unit modification during the synthesis process. We suggest that the resulting artefacts show similarities to those in disordered voices. To objectify diagnosis of those (pathological) voices, numerous acoustic markers capturing voice perturbation are utilized in clinical practice, yet, to our knowledge no studies exist for TTS signals. In the following, we present 4 traditional time-domain and 2 related cepstral perturbation measures which were evaluated on the basis of a TTS database for which formal listening tests were conducted.

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 where n denotes the n-th period of a (voiced) speech segment consisting of N periods [2]. The *Perturbation Factor* (PF) is then defined as

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

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

$$\mathrm{RAP} = \frac{\frac{1}{N-2}\sum_{n=1}^{N-2}\left|\frac{u(n)+u(n+1)+u(n+2)}{3}-u(n+1)\right|}{\frac{1}{N}\sum_{n=1}^{N}u(n)} \times 100\%. \qquad (2)$$

When $u(n)$ is equal to the fundamental period length $T_0(n)$, the above measures quantify jitter and are termed PPF and PRAP (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).

Perturbation in Cepstral Domain

According to a recent meta-analysis [4], the periodicity measure *Cepstral Peak Prominence* (CPP) [5] and its smoothed version (CPPs) have been proven to work best for overall voice-quality assessment when applied to continuous speech. CPP is defined as the height of the first rhamonic under a regression line through the cepstrum, used to normalize with respect to overall magnitude. Previous averaging of the cepstrum across time and across quefrency yields the CPPs. Both CPP and CPPs display how periodic a voiced signal is with respect to its excitation component, i.e how well-defined the harmonic configuration of the spectrum is.

TTS database

The described measures are applied to a TTS database from [6], based on six different concatenative TTS systems, where 17 participants rated 30 female and 30 male synthesized german sentences of about 12 seconds length each, following ITU-T Rec. P.85 (attribute-oriented listening-only test with absolute category rating (ACR) [7]). Eight attributes were considered: overall impression (MOS), listening effort (LSE), comprehension (CMP), articulation (ART), naturalness (NAT), prosody (PRO), continuity/fluency (CFL), and acceptance (ACC). Apart from ACC (binary scale), all attributes were rated on ACR-scales, ranging from 1 (bad) to 5 (excellent). The signals were preprocessed according to ITU-Rec. G.712 and normalized such that the average active speech level (ASL) matches −26 dBov. The sampling rate is 8 kHz.

Implementation

All measures are evaluated in voiced speech parts only for which a pitch estimation is provided using an autocorrelation-based algorithm [8] from the phonetics software Praat [9] and a subsequent epoch marking algorithm from Boersma as described in [10] (cross-correlation type). 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. This discrimination is reasonable in order to match a comparable blocklength for both genders. The frameshift is fixed to 10 ms. The smoothing with respect to CPPs is applied over time (3 blocks) and quefrency (10 bins). Perturbation measures, calculated per voiced segment (and per frame) are averaged across all voiced segments in order to yield per-signal values. These are then linearly regressed onto the naturalness scale, seperately for male and female stimuli. An overview of the system is given in figure 1.

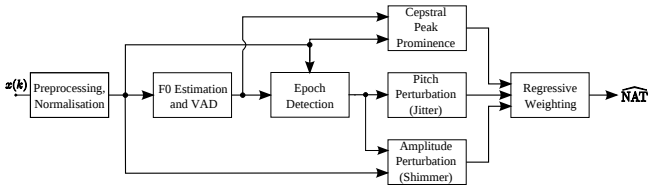

Figure 1: Overview of the quality estimator for a sampled TTS-signal $x(k)$.

Results and Discussion

For analysis purposes we calculate the Pearson (R) and the Spearman-rank (ρ) correlation coefficient between the subjective ratings and the perturbation measures. A high value R close to 1 indicates a prevalent linear relationship, whereas ρ is also sensitive to nonlinear dependencies. All values are given on a per-stimulus and per-synthesizer basis. Correlations are most pronounced across the naturalness scale, which was expected. The remaining correlation "distribution" across the attributes is similar for all perturbation measures (for both genders), showing medium values for MOS and the suprasegmental attributes (PRO, CFL, ACC) and lower values for the segmental attributes (LSE, CMP, ART). Hence, we only give the relationship between naturalness ratings and perturbation in table 1. Note that we keep minus

Table 1: Correlation between perturbation measures (PTB) and naturalness ratings, reported "per-stimulus/per-synthesizer".

PTB	CORRELATION			
	MALE		FEMALE	
	R	ρ	R	ρ
PPF	0.69/0.72	0.60/0.71	0.78/0.87	0.84/0.94
EPF	-0.04/0.03	0.07/0.09	0.64/0.77	0.58/0.77
PRAP	0.62/0.65	0.39/0.37	0.75/0.85	0.82/0.94
ERAP	-0.15/-0.17	0.02/-0.14	0.73/0.83	0.73/0.94
CPP	-0.36/-0.40	-0.24/-0.31	-0.49/-0.54	-0.35/-0.37
CPPs	-0.64/-0.69	-0.59/-0.77	-0.65/-0.70	-0.72/-0.89

signs to emphasize consistent inverse linear relationships with respect to the periodicity measures (CPP, CPPs) as opposed to the aperiodicity measures. Whereas Jitter (PPF and PRAP) shows a consistent behaviour for male and female signals, shimmer (EPF, ERAP) is only correlative for the female case. Generally, all perturbation measures perform better for the female stimuli. Considering the cepstral markers, CPP is less useful than CPPs which is in line with [4, 5], though CPP performs much weaker here, hence we drop CPP in favour of CPPs for our investigation. Most interestingly, perturbation shows a positive linear relationship with naturalness, in contrast to its deployment for disordered-voice diagnostics where higher perturbation typically indicates a less natural (healthy) voice. For a possible explanation we take into account that the database consists of diphone and unit-selection-type synthesis samples only, with the latter rated better in average due to the longer units incorporated. We presume that these units need to be less adapted during synthesis and the original degree of perturbation (from the inventory speaker) is thus better retained than in the diphone case. In this context, perturbation can be interpreted as an instrument for estimating the degree of residual naturalness induced from the inventory. After linear regression of PPF, PRAP, EPF, ERAP, and CPPs onto the naturalness scale, we yield a correlation of R = 0.84 and R = 0.86 for male and female stimuli, respectively (see figure 2), which reveals the usefulness of the presented approach. The consistency of the reported findings will be investigated further on broader databases in near future.

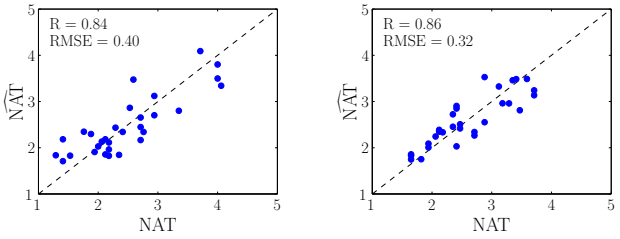

Figure 2: Auditive naturalness ratings (NAT) plotted against the estimated ratings ($\widehat{\text{NAT}}$) after regression for male (left) and female stimuli (right). The figures include Pearson's correlation coefficient (R) and the root-mean-square error (RMSE).

Acknowledgement

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

References

[1] D. H. Klatt and L. C. Klatt: Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.* **87(2)**, 820–857 (1990)

[2] D. Michaelis: Das Göttinger Heiserkeits-Diagramm. PhD thesis, Universität Göttingen (1999)

[3] N.B. Pinto and I. R. Titze: Unification of perturbation measures in speech signals. *J. Acoust. Soc. Am.* **87(3)**, 1278–1289 (1990)

[4] Y. Maryn et al.: Acoustic measurement of overall voice quality: A meta-analysis. *J. Acoust. Soc. Am.* **126(5)**, 2619–2634 (2009)

[5] J. Hillenbrand et al.: Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research* **39**, 311–321 (1996)

[6] K. Seget: Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren. Diplomarbeit, Christian-Albrechts-Universität zu Kiel, 2007

[7] ITU-T Rec. P.85, A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices. Int. Telecomm. Union, Geneva, 1994

[8] P. Boersma: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings* **17**, 97–110, University of Amsterdam, Netherlands, 1993

[9] P. Boersma and D. Weenik: Praat, software for speech analysis and synthesis, University of Amsterdam, 2005 `http://www.fon.hum.uva.nl/praat`

[10] M. Hagmüller and G. Kubin: Poincare pitch marks. *Speech Communication* **48**, 1650–1665 (2006)