Quality prediction for synthesized speech: Comparison of approaches

Sebastian Möller[1], Tiago H. Falk[2]

[1] *Deutsche Telekom Laboratories, TU Berlin, Germany, Email: sebastian.moeller@telekom.de*

[2] *Dept. of Electrical and Computer Engineering, Queen's University, Canada, Email: 2thf@queensu.ca*

Introduction

Text-To-Speech (TTS) technology has reached a level of maturity which seems to be sufficient for a number of telephony applications. In order to assess TTS quality, system developers need to carry out auditory tests where participants are asked to transcribe what they have heard or to rate certain aspects of the auditory event, see e.g. [1]. To overcome the temporal and financial effort involved in auditory testing, it is desirable to estimate the quality on the basis of speech signals, i.e. on the basis of instrumental measurements.

Several approaches for this purpose are described in the literature. Most of them address concatenative TTS systems where a natural speech corpus is available. For example, an average concatenative cost function has been used in [2] to assess the naturalness of concatenation-type synthesizers. The measure is derived from the input text and the natural speech corpus and is inversely proportional to overall quality – the higher the number of concatenations, the lower is the quality. Alternatively, signal based measures have been proposed which focus on computing spectral distances between the target synthesized speech signal and its original natural speech counterpart, see e.g. [3] and [4]. Such measures, however, are only useful if perceptual degradations are linked to concatenation effects and if a reference natural speech corpus is available; such requirements are usually not met in practice. To overcome such limitations, a measure which is based on the synthesized speech signal alone, without the need for a natural-speech reference, is required.

For natural speech, reference-free quality measurement algorithms are in common use, such as the standard algorithms ITU-T Rec. P.563 [5] and ANSI ANIQUE+ [6]. Such measures can in principle also be used for estimating TTS quality; however, they have been optimized for speech degraded by different transmission channel effects (noise, codec distortions, etc.), so they cannot be expected to provide adequate predictions for synthesized speech without further adaptation. In fact, own trials documented in [7] show that standard measures are well suited for estimating the transmission channel impact, but not the degradations coming with a synthesized source signal.

In this paper, we will compare two different reference-free quality measures on a TTS corpus: The first one is the standard ITU-T Rec. P.563 algorithm. It generates an artificial (high-quality) reference, compares the TTS signal to this reference, and combines the result with a noisiness, clipping and robotization detector in a parametric analysis to derive an overall quality estimate. The second approach is based on a Hidden Markov Model (HMM) trained on natural (male or female) speech; features derived from TTS signals are compared to the model for natural speech and quality is derived from the normalized log-likelihood between the reference models and the extracted features. The performance of both approaches is presented below, and proposals for improvements will be outlined in the final section.

Parametric Approach

The model which is recommended in ITU-T Rec. P.563 resulted from a competition held between 2002-2004 by ITU-T Study Group 12. It combines three approaches to obtain valid and reliable quality estimates [8], as can be seen in Figure 1. An in-depth description is given in the standard [5].

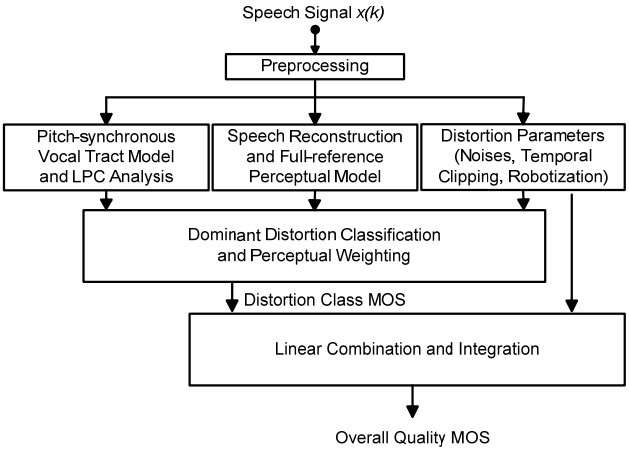

Figure 1: Sketch of the parametric approach according to [5].

Firstly, a vocal tract analysis is performed on the degraded speech signal, describing the (assumed) vocal tract as a series of tubes. Abnormal variations in the tubes' sections are considered as degradations. Secondly, a reference signal is reconstructed from a vocal tract synthesis, and this reference is compared to the degraded speech signal, in a signal-comparison approach similar to the one used in reference-based quality estimation like [9]. Thirdly, a number of parameters are determined which help to identify degradations and classify them according to one of six classes:

- Low static Signal-to-Noise Ratio, SNR
- Mutes and interruptions
- Low segmental SNR
- Unnatural voice – robotization
- Unnatural male voice

- Unnatural female voice

The overall quality is finally determined from the degradation measures weighted according to the dominant distortion identified in the speech signal, and a linear combination of additional parameters extracted from the distorted speech signal.

It should be noted that – due to its internal structure – the measure according to ITU-T Rec. P.563 does not only provide an estimate of the overall quality, but also of individual degradations. Thus, the model may also serve for the diagnosis of degradations introduced during the synthesis process.

HMM-based Approach

The signal processing steps involved in the computation of the proposed HMM-based quality measure are briefly depicted in Figure 2, and an in-depth description of the

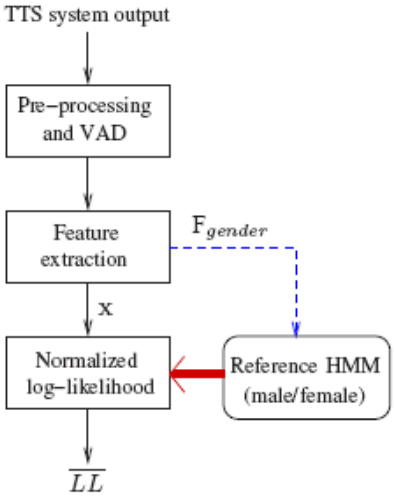

Figure 2: Sketch of the HMM-based approach according to [10].

calculation is given in [10].

The TTS signal is first pre-processed to match the characteristics of the signals used to develop the reference models. Voice Activity Detection (VAD) is then performed on the pre-processed signal to remove silence intervals longer than an empirically set value. The feature extraction module serves to compute perceptual and prosodic features. As pilot experiments have shown that performance can be improved if gender-dependent reference models are used, the prosodic features serve to identify talker gender. Lastly, perceptual features are assessed against offline-obtained reference HMMs of natural speech-feature behaviour via a normalized log-likelihood measure (LL). These references have been trained on the Kiel Corpus of German Read Speech, taking the "Siemens" and "Erlangen" sentence subsets, uttered by two male and two female speakers, as a basis. Per-gender files of approximately one hour and 15 minutes of active speech have been used to train a male and a female HMM reference model.

Comparison

In this section, we will compare the estimations of the two instrumental measures to the results of an auditory listening-only test. These test data have been kindly provided to us by LNS, Christian-Albrechts-Universität zu Kiel, Germany.

Test stimuli originate from six different TTS systems with German voices: 3 are commercial (AT&T, Proser, and Cepstral synthesizers), and 3 are from German academic institutions (Technical University Dresden, Technical University Berlin, and University of Bonn). The speech samples have been bandpass-filtered in the telephone range 300-3400 Hz, coded using the G.711 PCM speech codec, and level-normalized to -26dBov. Details on this experiment are described in [11].

The listening test mainly followed the recommendations in ITU-T Rec. P.85 [1] and was performed in a silent listening room at the Institute for Phonetics and Digital Speech Processing at Christian-Albrechts-Universität zu Kiel. Seventeen listeners (10 female, 7 male) participated in the test; all listeners were German students in the age range 20-26 years. Listeners were given a parallel task, and then asked to rate each stimulus on to eight rating scales related to the overall impression (resulting in a Mean Opinion Score, MOS), listening-effort (LSE), comprehension problems (CMP), articulation (ART), naturalness (NAT), prosody (PRO), continuity/fluency (CFL), and acceptance (ACC).

The performance of the instrumental measures is evaluated using the Pearson correlation coefficients between the output of the model (MOS or LL) and the auditory judgments, calculated on a per-sample basis. Table 1 reports the correlation coefficients for the male and the female samples, both considered separately or jointly (overall).

Table 1: Pearson correlation coefficients between auditory judgments and instrumental model estimations.

Auditory judgment	HMM-based (LL)			P.563		
	male	female	overall	male	female	overall
MOS	0.81	0.72	0.77	0.58	-0.05	0.24
LSE	0.72	0.64	0.65	0.50	0.02	0.20
CMP	0.70	0.45	0.54	0.42	-0.11	0.05
ART	0.74	0.47	0.55	0.53	-0.06	0.11
NAT	0.81	0.80	0.81	0.48	-0.06	0.24
PRO	0.54	0.72	0.61	0.28	-0.18	0.12
CFL	0.74	0.81	0.74	0.51	0.06	0.24
ACC	0.65	0.71	0.67	0.35	-0.10	0.15

The table shows that the proposed HMM log-likelihood measure LL correlates well with several quality dimensions, in particular with MOS, NAT, and CFL. Interestingly, LL computed for male speech obtains considerably higher correlation values than the one computed for female speech, for quality dimensions CMP and ART. In turn, the correlation is higher with female data for the dimension PRO. Relative to the measure of ITU-T Rec. P.563, substantially higher correlation values are attained with the proposed LL measure in all cases. Note also that poor correlations are attained with P.563 for female synthesized speech; such intriguing behaviour has also been reported in

[12] for synthesized speech transmitted over noisy telephone channels.

In order to get an impression of where the lower performance of the P.563 model stems from, we analyzed the behaviour of some internal parameters of this model. First, we identified the dominant distortion class detected by the model. We observed that the P.563 model selects distortion classes "unnatural male" 33.33% of the time, "unnatural female" 31.66%, "mutes/interruptions" 33.33%, and "robotization" 1.66% percent of the time, respectively.

The "unnatural male" class is defined by a sufficiently high SNR, no apparent mutes/interruptions/sharp declines, no excessive periodicity of the voices signal parts, and a pitch frequency lower than 160 Hz. The corresponding "unnatural female" class is defined by the same characteristics, but a higher fundamental frequency. Apparently, the degradations of our synthesized speech signals are not captured by the parameters used for these dominant distortion classes. Still, about one third of the dominant distortions are classified as "mutes/interruptions", indicating that this type of degradation might be responsible for a part of the audible effects.

For the dominant distortion, the P.563 model calculates a basic "distortion class MOS". The Pearson correlation coefficient between this distortion class MOS and the auditory MOS rating is only 0.02 overall (0.42 for the male and -0.29 for the female samples), i.e. far lower than the aforementioned performance obtained with the final MOS estimated by the P.563 model. Such lower performance suggests that the eleven features used in the final MOS mapping are more important for synthesized speech quality measurement than the features used to compute the distortion class MOS value.

Table 2: Correlations between internal parameters of the P.563 model and auditory MOS scores.

Female		Male	
Parameter	*R*	Parameter	*R*
VTP VAD Overlap	-0.58	Art Average	-0.64
Basic voice quality	-0.63	Pitch cross power	-0.55
Basic voice qual. asym.	-0.74	Basic voice quality	0.53
LPC curtosis	0.71	Cepstral curtosis	0.65
Spectral level dev.	-0.64	*Cepstral skew*	0.69
Relative noise floor	0.60	High frequ. variation	0.63
		Global backgr. noise	-0.73
		Spectral level range	0.59
		Sharp declines	-0.56

To further investigate this claim, correlations between each internal feature computed by P.563 and the auditory MOS scores are computed. Table 2 reports all parameters that attain Pearson correlation values $|R| > 0.5$ for either the female or the male TTS samples. Although a considerable number of parameters match this criterion, only three of them (set in *italics*) are part of the final integration function adjusting the distortion class MOS to the final MOS score; the other parameters (set in normal font) are internal parameters used to calculate the distortion class MOS. This shows that not only the final integration function has to be elaborated, but also the definition of the distortion classes itself, and the underlying parameters. Still, the substantially higher correlations attained with some internal parameters suggest that improved parameter combinations, distortion classes and mapping functions may help to increase the prediction performance.

It has to be noted the correlations for the parameters "basic voice quality" show an inverse sign for the female and the male speech files. Similar observations have also been made in [12] for other parameters and test databases. Apparently, it is advantageous if the analysis is carried out separately for male and female speech files. This finding coincides with the observations already reported for the HMM-based approach, where separate reference models for male and female speech files also improve prediction performance.

Conclusions

The reported results show that an instrumental prediction of TTS quality is feasible. Although the performance is not yet comparable to the one observed for degraded natural speech (where correlations above 0.90 are not rare), the initial results make us confident that such approaches may be adequate to

- obtain an initial impression of the overall performance of a TTS system,
- rank different TTS systems with respect to their quality,
- perform a diagnosis of the most disturbing characteristics of a TTS system, and
- select TTS systems from a pool for a more thorough auditory test.

Still, we are convinced that instrumental models will not be able to completely replace auditory tests. In fact, such models are only able to recapitulate judgments for degradation they have been trained on; each time a new TTS system is developed which potentially introduces new types of degradations, the validity and reliability of such a prediction algorithm has to be tested anew.

In order to increase the prediction performance, a more in-depth analysis with a larger pool of test data is needed. Our current plan is to make use of the Blizzard campaigns organized each year for comparing the latest developments of TTS systems. Initial results have been presented in [13], and they show that a prediction is also feasible with unknown test data.

In addition to the prediction of overall quality, we would like to address the prediction of auditory quality dimensions. The dimensions addressed in the described test may serve as an example, but they are probably not orthogonal. Such orthogonal dimensions may be identified e.g. through multidimensional scaling experiments, as it has been done for transmitted natural speech [14]. If we are able to predict perceptual dimensions in a reliable way, we think that it will be possible to also obtain overall quality estimations by combining several perceptual dimension indicators. Such a dimension-based approach may be less vulnerable to test- and system specific effects, provided that all relevant

perceptual dimensions are covered, and it may provide diagnostic information which is valuable for TTS system developers.

Acknowledgment

The authors would like to thank the Institute of Circuit and System Theory (LNS), Christian Albrechts-Universität zu Kiel, for kindly providing the test data.

References

[1] ITU-T Rec. P.85, Method for subjective performance assessment of the quality of speech voice output devices. International Telecommunication Union, Geneva, 1994.

[2] Chu, M., Peng, H., An objective measure for estimating MOS of synthesized speech. In: Proc. European Conf. Speech Communications and Technology, Aalborg, 2001, 2087-2090.

[3] Cernak, M., Rusko, M., An evaluation of synthetic speech using the PESQ measure. In: Proc. Forum Acusticum, Budapest, 2005, 2725-2728.

[4] Vepa, J., King, S., Taylor, P., Objective distance measures for spectral discontinuities in concatenative speech synthesis. In: Proc. Intl. Conf. Spoken Language Proc., Sept. 2002, 2605-2608.

[5] ITU-T Rec. P.563, Single ended method for objective speech quality assessment in narrowband telephony applications. International Telecommunication Union, Geneva, 2004.

[6] ATIS-PP-0100005.2006, Auditory non-intrusive quality estimation plus (ANIQUE+): Perceptual model for non-intrusive estimation of narrowband speech quality. American National Standards Institute, 2006.

[7] Möller, S., Kim, D.-S., Malfait, L., Estimating the quality of synthesized and natural speech transmitted through telephone networks using single-ended prediction models, Acta Acustica united with Acustica **94** (2008), 21-31.

[8] Malfait, L., Berger, J., Kastner, M., P.563 – The ITU-T standard for single-ended speech quality assessment. IEEE Trans. Audio, Speech, Lang. Process. **14** (2006) 1924-1934.

[9] ITU-T Rec. P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. International Telecommunication Union, Geneva, 2001.

[10] Falk, T. H., Möller, S., Towards signal-based instrumental quality diagnosis for text-to-speech systems. IEEE Signal Processing Letters **15** (2008), 781-784.

[11] Seget, K., Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren (Studies on the auditory quality of voice synthesis methods). Bachelor thesis, LNS, Christian-Albrechts-Universität zu Kiel, 2007.

[12] ITU-T Contr. COM 12-180, Single-ended quality estimation of synthesized speech: Analysis of the Rec. P.563 internal signal processing. Federal Republic of Germany (Authors: S. Möller, T.H. Falk), ITU-T SG12 Meeting, 22-29 May, Geneva, 2008.

[13] Falk, T. H., Möller, S., Karaiskos, V., King, S., Improving instrumental quality prediction performance for the Blizzard Challenge. In: Proc. Blizzard Challenge Workshop, Brisbane, 2008, 6 pages.

[14] Wältermann, M., Möller, S., Raake, M., Scholz, K., Huo, L., Heute, U., An instrumental measure for end-to-end speech transmission quality based on perceptual dimensions: Framework and realization. In: Proc. Interspeech 2008 incorporating SST 2008, Brisbane, 22-26 Sept. 2008, ISCA, 61-64.