Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems

Tiago H. Falk, *Student Member, IEEE*, and Sebastian Möller

***Abstract*—In this letter, the first steps toward the development of a signal-based instrumental quality measure for text-to-speech (TTS) systems are described. Hidden Markov models (HMM), trained on naturally-produced speech, serve as artificial text- and speaker-independent reference models against which synthesized speech signals are assessed. A normalized log-likelihood measure, computed between perceptual features extracted from synthesized speech and a gender-dependent HMM reference model, is proposed and shown to be a reliable parameter for multidimensional TTS quality diagnosis. Experiments with subjectively scored synthesized speech data show that the proposed measure attains promising estimation performance for quality dimensions labeled *overall impression, listening effort, naturalness, continuity/fluency*, and *acceptance*.**

***Index Terms*—Hidden Markov model, multidimensional quality diagnosis, quality prediction, synthesized speech, text-to-speech (TTS).**

I. Introduction

TEXT-TO-SPEECH (TTS) synthesis, as the name suggests, attempts to convert arbitrary input text into intelligible and naturally-sounding speech. Earlier applications of TTS systems served mostly as an aid to the visually impaired. Today, TTS systems are also being applied in e-mail and short message service readers, automated directory assistance, foreign language education, and assistive and augmentative communications. As new applications emerge, the need to deliver high-quality synthesized speech increases. As such, the demand on methods to evaluate the quality of TTS systems has also risen.

Evaluation of synthesized speech, however, is not an easy task as various quality dimensions can be assessed (e.g., naturalness, intelligibility). Commonly, multidimensional subjective listening quality tests, such as the one described in the International Telecommunication Union Recommendation ITU-T Rec. P.85 [1], are used. Major drawbacks of subjective assessment, however, are the high costs and intense labor associated with running the tests. For applications such as TTS system tuning, several tests may be required throughout the development process; in such case, subjective assessment is not feasible and instrumental quality prediction is needed.

To date, there is no universally accepted signal-based instrumental quality measure for synthesized speech. Most measures are for corpus-based concatenative TTS systems where the *natural speech* corpus is available. In [2], an average concatenative cost function is used to assess the naturalness of concatenation-type synthesizers. The measure is derived from the input text and 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 and focus mainly on computing spectral distances between the target synthesized speech signal and its original natural speech counterpart (e.g., see [3]–[5] and references therein). 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 not always met in practice. To overcome such limitations, a "reference-free" measure is required.

For *natural* speech, reference-free (also termed non-intrusive) quality measurement algorithms have been proposed, such as standard algorithms ITU-T Rec. P.563 [6] and ANSI ANIQUE+ [7]. To the best of our knowledge, a signal-based reference-free quality measure for *synthesized* speech has yet to emerge. Recently, the aforementioned reference-free standard algorithms, developed for natural speech, were tested on synthesized speech transmitted over different telephone channels [8]. While the measures were shown to estimate the effects of the transmission channel, poor estimation of source speech quality was attained, signaling the need for a more accurate quality measure for synthesized speech.

In this letter, the first steps towards devising a general-purpose signal-based *reference-free* measure for TTS system quality diagnosis are described. In particular, hidden Markov models are used to devise text- and speaker-independent artificial reference models of naturally-produced speech-feature behavior. Perceptual features, extracted from synthesized speech, are then assessed against gender-dependent reference models by means of a normalized log-likelihood measure. The degree of "consistency" with the reference models is proposed as a measure for multidimensional quality diagnosis.

The remainder of this letter is organized as follows. Section II provides an overview of subjective diagnosis of TTS system quality. Section III describes the signal processing steps needed to compute the proposed quality measure. Section IV describes the databases used in our tests and reports the experimental results. Conclusions are presented in Section V.

II. Subjective Diagnosis of TTS System Quality

Speech quality is the result of a subjective perception-and-judgment process, during which a listener compares the perceptual event (speech signal heard) to an internal reference of what

Manuscript received July 03, 2008; revised August 20, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philip C. Loizou.

T. H. Falk is with the Department of Electrical and Computer Engineering, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: falkt@ee.queensu.ca).

S. Möller is with Deutsche Telekom Labs, Berlin Institute of Technology, Berlin D-10587, Germany (e-mail: sebastian.moeller@telekom.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2008.2006709

1070-9908/$25.00 © 2008 IEEE

is judged to be "good quality." The result of this comparison is usually quantified using one or multiple scales. The scale most commonly used is the five-point absolute category rating (ACR) scale where a rating of 1 corresponds to bad speech quality and a rating of 5 to excellent speech quality. With the ACR scale, the average of the listener scores constitutes the mean opinion score (MOS). For synthesized speech, multidimensional quality tests, such as the one described in [1], have been proposed. In the test, listeners are asked to rate the signal using eight quality dimensions labeled: *overall impression, listening effort, comprehension problems, articulation, pronunciation, speaking rate, voice pleasantness*, and *acceptance*. For the first seven listed dimensions, a five-point scale is used (see [1] for more detail). For the acceptance dimension, a two-point scale (yes/no) is used, and results are reported as a percentage acceptance value.

During the test, subjects are presented with each speech file twice. In the first presentation, subjects are asked to solve a secondary task such as answer specific questions about information contained in the file (e.g., bus date/time of departure). Subjects are then asked to judge the quality of the speech signal based on the aforementioned quality dimensions. The intent of providing a secondary task is to direct the listeners' attention to the content of the speech signal, and not on its surface form alone, so as to improve listener judgement of quality dimensions such as comprehension problems and listening effort. Despite being the most valid quality measurement methodology, subjective tests are expensive and time consuming, thus not suitable for applications such as system evaluation throughout the development cycle, where several system updates may need to be assessed. For such scenarios, instrumental quality measurement is needed.

III. Proposed HMM-Based Quality Measure

The signal processing steps involved in the computation of the proposed HMM-based quality measure are depicted in Fig. 1. Preprocessing is first performed to match the characteristics of the signals used to develop the reference models. Voice activity detection (VAD) is then performed on the preprocessed speech signal to remove silence intervals longer than an empirically set value. The feature extraction module serves to compute perceptual and prosodic features; the latter are used to identify talker gender. Pilot experiments have suggested that improved performance is attained if gender-dependent reference models are used. Lastly, perceptual features are assessed against offline-obtained reference hidden Markov models of natural speech-feature behavior via a normalized log-likelihood measure. A detailed description of the signal processing steps is given in the subsections to follow.

A. Preprocessing, VAD, and Feature Extraction

In order to match the characteristics of the signals used to train the reference models, preprocessing is applied to the TTS system output. Representative preprocessing steps can include resampling, filtering, and/or signal level normalization. In our experiments, preprocessing consists of bandpass filtering according to [9], downsampling to 8 kHz, and level normalization to -26 dBov (overload) using the P.56 speech-level meter [10]. Moreover, since we are interested in measuring the quality of the output of a TTS system, only active speech segments are analyzed. In our experiments, a simple energy thresholding VAD algorithm is used to remove silence intervals longer than 75 ms; such duration is empirically chosen as to avoid "artificial" discontinuities introduced by possible VAD errors.

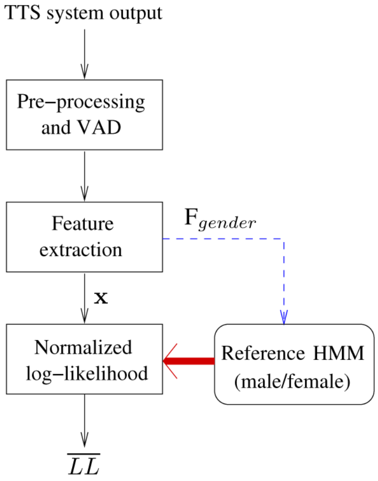

Fig. 1. Signal processing steps involved in the computation of the proposed HMM-based quality measure. Separate hidden Markov reference models of natural speech-feature behavior are used for male and female speech.

Perceptual features are then computed from active speech; features include 12th-order mel-frequency cepstral coefficients (MFCC). The notation $\mathbf{c}_m = \{c_{0,m}, \ldots, c_{12,m}\}$ is used to represent MFCC computed for speech frame m. In our experiments, MFCCs are computed using 25-ms windows and 10-ms shifts. The zeroth order cepstral coefficient $c_{0,m}$ is used as a log-energy measure. A basic assumption used in this study is that, for natural speech, abrupt changes in signal energy do not occur. Such discontinuities, however, can occur in, e.g., speech produced by a concatenative TTS system. In order to quantify signal-energy dynamics, we compute the zeroth delta-cepstral coefficient $\Delta c_{0,m}$, which has been shown useful for temporal discontinuity detection [11]. Feature $\Delta c_{0,m}$ is appended to $\mathbf{c}_m$ to form $\mathbf{x}_m = [\Delta c_{0,m}, \mathbf{c}_m]$. In Fig. 1, $\mathbf{x}$ constitutes features computed for the M active frames in the synthesized speech signal, i.e., $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^{M}$.

Lastly, the fundamental frequency $F0$ is computed with the pitch tracking algorithm described in [12]. $F0$, averaged over all voiced speech frames, is used to identify talker gender. In pilot experiments, it has been observed that improved quality measurement performance is attained if gender-dependent reference models are used. Motivated by the work described in [6], $F0 = 160$ Hz is used as a threshold to distinguish between male and female voices. A flag indicating talker gender, represented by $\mathrm{F}_{\mathrm{gender}}$ in Fig. 1, is used to indicate which HMM reference model to use.

B. HMM Reference Models and Log-Likelihood Computation

Speech temporal dynamics provides important information for the measurement of synthesized speech quality and naturalness. As such, we propose to use hidden Markov reference models trained on naturally-produced speech. The spectral-temporal information captured by the HMM can be used to quantify differences between, e.g., natural word endings and abnormal signal interruptions that may occur with synthesized speech. Reference models are obtained using the perceptual features $\mathbf{x}$ described in Section III-A. Features are extracted from natural speech (described in Section IV-A1) and two reference models are designed, one for male and one for female speech data.

Hidden Markov models with eight states are used and the output distribution of each state consists of a Gaussian mixture density with 16 diagonal-covariance Gaussian components. Model parameters, such as state transition probabilities, initial state probabilities, and output distribution parameters, are computed using the expectation-maximization algorithm summarized in [13]. Perceptual features, extracted from the synthesized signal under test, are then assessed against the reference models via the log-likelihood measure. Log-likelihood values are computed using the so-called forward-backward procedure described in [13]. Normalization is performed based on the number of active-speech frames M in the signal under test; normalized log-likelihood is referred to as $\overline{LL}$ in Fig. 1.

IV. Experiments

In this section, a description of the databases used in our tests, as well as the experiment results, are reported.

A. Database Description

A description of the naturally-produced and synthesized speech databases used in our experiments is given in the subsections to follow. Natural speech is used to train gender-dependent HMM reference models and synthesized speech to assess the performance of the proposed quality measure.

1) Natural Speech—Training Data: In order to develop reference models of natural speech-feature behavior, the Kiel Corpus of German read speech is used. Files from the "Siemens" and "Erlangen" sentence subsets, uttered by two male and two female speakers, are used. Visual inspection of spectrograms and pitch contours was used to select speakers with spectral-temporal characteristics different from those in the synthesized speech database. The files are downsampled to 8 kHz, bandpass filtered according to [9], level normalized to −26 dBov, and VAD-processed. Per-gender files are concatenated to produce approximately 1 h and 15 min of active speech to train the male and female HMM reference models. It is emphasized that the sentences uttered in the training speech dataset differ from the text used to generate the synthesized speech material.

2) Synthesized Speech—Test Data: The synthesized speech database used in our experiments contains speech material from six "off-the-shelf" TTS systems. Three are commercial systems (AT&T, MBROLA-based Proser, and Cepstral), and three are from German academic institutions (TU Dresden, TU Berlin, and University of Bonn). Speech material is produced from the system online demonstration tool; thus, in the case of corpus-based TTS systems, access is *not* available to the natural speech corpus. A total of ten speech samples are generated per TTS system, half for male speakers and half for female. The synthesized speech samples have an average duration of 11 s and consist of two utterances separated by a silence interval of approximately 2 s. Speech samples were bandpass-filtered according to [9] and level-normalized to −26 dBov prior to listener presentation.

The listening test closely 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-University of Kiel [14]. Seventeen listeners (ten female, seven male) participated in the test; all were German students and the age ranged from 20–26. Listeners were given a parallel task and asked to rate the synthesized speech signals using eight quality scales. Of the eight scales used, only five are described in ITU-T Rec. P.85. Labels of the eight scales used include: overall impression (MOS), listening effort (LSE), comprehension problems (CMP), articulation (ART), naturalness (NAT), prosody similarity with natural speech (PRO), continuity/fluency (CFL), and acceptance (ACC). Table I reports the rating scales for dimensions NAT, PRO, and CFL; scales for the five remaining dimensions are described in [1].

TABLE I
Rating Scales Used in the Listening Test Not Described in [1]. Original Wordings in German are Reported in [14]

Rating	NAT	PRO	CFL
5	Very natural	Very similar	Very fluent
4	Natural	Similar	Fluent
3	Neutral	Somewhat similar	Neutral
2	Unnatural	Dissimilar	Discontinuous
1	Very unnatural	Very dissimilar	Very discontinuous

TABLE II
Performance Comparison Between $\overline{LL}$ and ITU-T Rec. P.563 on Eight Synthesized Speech Quality Dimensions

Quality	Proposed $\overline{LL}$			ITU-T P.563		
dimension	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

B. Experiment Results

To test the performance of the proposed quality measure, Pearson correlation coefficients attained between $\overline{LL}$ and the various quality dimensions are computed. Table II reports "per speech sample" correlation coefficients attained for the eight quality dimensions for male and female speech data, considered either separately or jointly (overall). For comparison purposes, correlation coefficients attained with the state-of-art ITU-T Rec. P.563 algorithm are also reported. It is emphasized, however, that synthesized speech does *not* fall within the recommended scope of the standard P.563 algorithm. Unfortunately, no other signal-based reference-free measures are available for comparison.

As observed from the table, the proposed HMM log-likelihood measure correlates well with several quality dimensions, in particular with MOS, NAT, and CFL. Interestingly, $\overline{LL}$ computed for male speech obtains considerably higher correlation values, relative to female speech, for quality dimensions CMP and ART. In turn, higher correlation is attained with female data for dimension PRO. Relative to P.563, substantially higher correlation values are attained with the proposed $\overline{LL}$ measure.

TABLE III
PERFORMANCE COMPARISON BETWEEN $\overline{LL}$ AND ITU-T REC. P.563 AFTER THIRD-ORDER (GENDER-DEPENDENT) POLYNOMIAL REGRESSION

Quality	Proposed $\overline{LL}$		ITU-T P.563	
dimension	Male	Female	Male	Female
MOS	0.83	0.74	0.65	0.05
LSE	0.74	0.70	0.59	0.07
CMP	0.72	0.56	0.48	0.02
ART	0.78	0.57	0.62	0.01
NAT	0.84	0.83	0.59	0.20
PRO	0.61	0.72	0.39	0.07
CFL	0.79	0.82	0.61	0.07
ACC	0.70	0.73	0.47	0.03

Note also that poor correlations are attained with P.563 for female synthesized speech; such intriguing behavior has also been reported in [15] for synthesized speech transmitted over noisy telephone channels.

Furthermore, the work described in [16] suggests cross-gender differences in the subjective perception of synthesized speech quality. In an attempt to compensate for such listener rating "biases," a gender-dependent monotonic polynomial mapping function is applied between $\overline{LL}$ and the subjective quality scores. Monotonic mappings perform scale adjustments but do not alter the ranking of the estimated scores. Table III reports correlation coefficients attained *after* third-order polynomial regression. As can be seen, a slight improvement in performance is attained after regression; for P.563 predictions, poor correlations remain for female speech.

Ultimately, the aim in instrumental measurement is to develop a measure that ranks similarly with subjective quality ratings. To this end, Spearman rank correlation is computed and used as an additional figure of merit. Spearman correlation is computed in a manner similar to Pearson correlation, except original data values are replaced by the *ranks* of the data values. Due to space constraints, Spearman correlation coefficients ρ_S are only reported for quality dimension MOS. On our data, the proposed $\overline{LL}$ measure attains $\rho_S = 0.76$ and $\rho_S = 0.70$ for male and female data, respectively. For P.563, $\rho_S = 0.57$ and $\rho_S = 0.03$ are obtained, respectively.

C. Discussion

While the proposed measure is shown to correlate well with several quality dimensions, it is inferred that further performance gains can be attained if additional features are used in combination with $\overline{LL}$. Representative features can include the mean cepstral deviation, proposed in [11] as a measure of spectral flatness, which has also been shown useful for spoken dialogue system evaluation [17]. On our data, mean cesptral deviation attains correlation values of -0.64, -0.62, and -0.61 with LSE, CMP, and NAT, respectively (for female speech). Moreover, a sharp decline measure, similar to the one described in [6], is shown to attain correlation values of -0.56, -0.57, and -0.62 with CMP, PRO, and CFL, respectively (for male speech). Feature combination, however, requires access to multiple subjectively scored speech databases in order to optimize feature weights and hence is left for a future study.

V. CONCLUSION

The first steps towards the development of a general-purpose signal-based instrumental quality measure for text-to-speech systems are described. The measure, based on hidden Markov reference models of naturally-produced speech, is shown to attain promising results on a multidimensional quality prediction test for both male and female synthesized speech.

ACKNOWLEDGMENT

The authors would like to thank L. Huo, K. Seget, and U. Heute, all from Christian-Albrechts-University of Kiel, Germany, for obtaining the listening test results.

REFERENCES

[1] *Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*, ITU-T Rec. P.85, 1994, Int. Telecom. Union.

[2] M. Chu and H. Peng, "An objective measure for estimating MOS of synthesized speech," in *Proc. Eur. Conf. Speech Communications and Technology*, 2001, pp. 2087–2090.

[3] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. Eur. Congr. Acoustics*, 2005, pp. 2725–2728.

[4] J. Vepa, S. King, and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," in *Proc. Int. Conf. Spoken Language Processing*, Sep. 2002, pp. 2605–2608.

[5] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Jun. 2001, pp. 837–840.

[6] *Single Ended Method for Objective Speech Quality Assessment in Narrowband Telephony Applications*, ITU-T Rec. P.563, 2004, Int. Telecom. Union.

[7] *Auditory Non-Intrusive Quality Estimation Plus (ANIQUE+): Perceptual Model for Non-Intrusive Estimation of Narrowband Speech Quality*, ATIS-PP-0100005.2006, 2006, Amer. Nat. Standards Inst..

[8] S. Möller, D.-S. Kim, and L. Malfait, "Estimating the quality of synthesized and natural speech transmitted through telephone networks using single-ended prediction models," *Acta Acustica United With Acustica*, vol. 94, pp. 21–31, 2008.

[9] *Transmission Performance Characteristics of Pulse Code Modulation Channels*, ITU-T Rec. G.712, 2001, Int. Telecom. Union.

[10] *Objective Measurement of Active Speech Level*, ITU-T Rec. P.56, 1993, Int. Telecom. Union.

[11] T. H. Falk and W.-Y. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[12] "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Synthesis.*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier Science, 1995, pp. 495–518.

[13] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[14] K. Seget, "Untersuchungen Zur Auditiven Qualität von Sprachsyntheseverfahren (Study of perceptual quality of text-to-speech systems)," Bachelor thesis, Christian-Albrechts-Univ. Kiel, , Jul. 2007.

[15] S. Möller and T. H. Falk, Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing, 2008, ITU-T Contribution COM 12-180, Int. Telecom. Union.

[16] J. Mullennix, S. Stern, S. Wilson, and C. Dyson, "Social perception of male and female computer synthesized speech," *Comput. Human Beh.*, vol. 19, pp. 407–424, 2003.

[17] S. Möller, K.-P. Engelbrecht, M. Pucher, P. Frölich, L. Huo, U. Heute, and F. Oberle, "TIDE: A testbed for interactive spoken dialogue system evaluation," in *Proc. Int. Conf. Speech and Computers*, Oct. 2007.