Prediction of Perceived Sound Quality of Synthetic Speech

Dong-Yan Huang

Department of Signal Processing
*Institute for Infocomm Research/A*STAR*
1 Fusionopolis Way, # 21-01 Connexis (South Tower), Singapore 138632
{huang@i2r.a-star.edu.sg}

***Abstract*—This paper investigates the performance of objective speech and audio quality measures for the prediction of perceived sound quality of synthetic speech. A number of existing quality measures have been applied to synthetic speech generated by different speech synthesizers such like LP synthesizer, HSM synthesizer, STRAIGHT synthesizer and several HMM based text-to-speech synthesis systems. The subjective quality rating were obtained using the ITU-T P.85 methodology designed to evaluate the quality of synthetic speech along three dimension: speech naturalness, speech similarity, and overall quality. The correlation of several quality measures with these three subjective rating scales were evaluated among normal subjects. This paper reports the correlations of five objective measures with these three subjective measures and point out the research direction in the future.**

I. Introduction

Speech synthesis is the artificial production of human speech. A system developed for this purpose is called speech synthesizer, and can be implemented in software or hardware. Some systems can converts normal language text into speech (text-to-speech (TTS)); other systems can render symbolic linguistic representation like phonetic transcriptions into speech [1].

Synthesized speech can be created by concatenating of recorded speech that are stored in a database or a synthesizer incorporated a model of the vocal tract and other human voice characteristics. The quality of a speech synthesizer is assessed by its similarity to the human voice and by its ability to be understood. However, the consistent evaluation of speech synthesis systems may be difficult because of a lack of universally agreed objective evaluation criteria. Different organizations often use different speech data. Recently, some researchers have started to evaluate speech synthesis systems using a common speech dataset [2]. The perceived quality of synthetic speech will depend on a number of variables, including the parameter setting of algorithms, the production technique and the facilities used to replay the speech, listeners, type of sound, as well as highly individual, subjective factors like personal experiences, expectations and preferences.

The most popular method for evaluating the quality of synthesized speech is through subjective listening test. However, the output of the subjective evaluation of the quality of synthesized speech cannot simply be measured according to its accuracy. In general, mean opinion scores (MOS) are used for gathering subjective judgments of speech quality. MOS are obtained by analysis of untrained normal listeners ratings of stimuli along a scale. However, a number of studies have found that raters are often influenced by dimensions of the signal other than those they have been asked to rate. For example, listeners judgements of intonation naturalness have been shown to be affected by segmental quality [3], [4], while intonation appropriateness has been found to impact on perceived segmental quality [4]. Fact from speech perception and general auditory perception shows that listeners' hierarchies of weighting can differ depending on the segmental and acoustic context of the stimuli (e.g., speech versus non-speech, natural speech versus synthetic speech, first language versus second language,etc.) [5], [6], [7], [8]. A few years ago, multidimensional scaling (MDS) has been proposed as a tool for identifying the main acoustic dimensions to which listeners attend when rating synthetic speech [9]. The article addressed questions whether MDS can be introduced as a standard tool for large scale evaluation of the quality of synthesized speech and showed that firstly, Simple Euclidean MDS is an appropriate representation for individual listeners' judgement of speech, which is the precondition for large-scale evaluation, and secondly, MDS offers the same information content as MOS through the elaborate data collection, thirdly given a representative subset, smaller-scale experiments can predict larger-scale experiments' results .

As we know that this most popular approach follows a data-driven, i.e. statistical model approach, which requires a large base of empirical data containing subjective quality rating of different types of sounds, processed with a variety of a speech synthesizer setting, obtained from many listeners. It is costly and time/resource consuming. Therefore, an objective measure for speech synthesizer evaluation is the desirable research goal of this paper. Many objective speech quality measures have been proposed in the past to predict the subjective quality of speech [10]. Most of these measures, however, were dedicated for the purpose of evaluating the distortions introduced by speech codecs and/or communication channels [11], [12], [13], [14], [15], [25]. Later, some of measures and several new composite measures were evaluated for predicting the quality of speech enhanced by noise suppression algorithms [27]. Although the different types of distortions introduced by waveform and operations in speech synthesizer are different from those introduced by different speech enhancement algorithms, it has some similarity of speech codecs using analysis-by-synthesis methods. Hence, it is still worthy of investigating the possibility of using objective measures for predicting the quality of synthesized speech.

The types of distortion introduced by waveform and operations in synthesizers can be broadly classified into two categories: the distortions that affect the speech signal itself (called speech distortion which affects the naturalness of synthesized speech and original speech) and the distortion that affects the background noise (called noise distortion which affects the similarity of synthesized speech). According to our experiments, of these two types of distortion, listeners seem to be influenced the most by the speech distortion when making judgments of overall quality.

Compared to the speech coding literature, only a small number of studies examined the correlation between the objective measures and the subjective quality of TTS [16], [17], [18]. Mariniak [19] proposed to extract perception-based features from the synthesized speech material and to compare them to features extracted from (other) natural speakers; the distance between both could be an indication of the speech quality. To our knowledge, this approach was never implemented by Mariniak, but it has recently been taken up in [20], using Mel- Frequency Cepstral Coefficients (MFCCs) as features and a Hidden Markov Model (HMM) with Gaussian Mixture densities for a temporal-spectral comparison of features. It led to very promising results on an initial test database, with correlations between 0.54 and 0.81 for different quality dimensions collected in the auditory test. Another approach is to extract parameters from the speech signal which are related to degradations typically expected for TTS. Also this approach is motivated by quality prediction models for transmitted natural speech, namely the single ended model given in ITU-T Rec. P.563 [21]. This model first generates a clean speech reference from the degraded one, then calculates a perceptually-motivated distance between the degraded and the clean speech signal, further extracts a large number of parameters related to typical transmission channel degradations, and combines the perceptually weighted distance and the parameters to an estimation of overall speech quality. Applying this model to synthesized speech [22], the results were not as promising as those obtained with the HMM-based approach, but the parameters have not yet been optimized for synthesized speech. A comparison of different such single-ended speech quality models described in [22] shows that the P.563 model might not be the most appropriate one. In addition, considerable differences have been detected between the performances for male vs. female voices [23].

However, most studies reported correlation of objective measures with only overall quality of TTS systems using the objective measures based on some perceptual model [25]. These sound quality models compare ”internal representations”, computed by a psychoacoustic model, of a test and a reference sound signal [25]. Detected differences between internal representations are interpreted as quality degradations of the test signal with respect to the reference signal. Hence, these comparison-based models depend on the availability of a reference signal that represents the optimum, or desired, sound quality. This requirement is met in the evaluation of lossy signal processing systems, such as low-bitrate speech and audio codecs, speech/audio synthesizer, where the unprocessed, original signal serves as a reference. However, there is no ideal reference for the evaluation of transformed speech such like gender/age transformed speech, emotionally transformed speech,..etc. Recently, some new non-intrusive objective measures have appeared which do not need the acoustical reference [29].

To our knowledge, no comprehensive study was done to assess the correlation of existing objective measures with distortions (background and speech) present in synthesized speech and with the overall quality of synthesized speech. Since different speech synthesizers introduce different types of signal/background distortion, it is necessary to include various classes of speech synthesizers in such an evaluation. The main objective of the study is to report on the evaluation of conventional as well as several new composite objective measures that could be used to predict overall speech quality and speech/noise distortions introduced by TTS and speech model-based synthesizers. To that end, we make use of an existing subjective database for the evaluation of speech synthesizers. The subjective quality rating were obtained using the ITU-T P.35 methodology designed to evaluate the speech quality along three dimensions: signal distortion, noise distortion, and overall quality [26].

The paper is organized as follows. Section III describes the speech corpus and the subjective speech quality evaluation protocols. Section IV introduces the objective measurements and their application for sound quality prediction. Section V presents the resulting correlation coefficients. Section VI discusses the approach and results. Finally, Section VII gives the conclusion.

II. Method

For the evaluation of several objective measures, we use a subset of entries in the Blizzard challenge 2008, test set A were selected to conduct more extensive multiple dimension scaling analysis. following method to evaluate the objective measures for prediction of sound quality of synthetic speech. The length of the sentences ranges from 1.38 to 3.4 seconds, and from 8 to 15 syllables (I). This corpus was used in a comprehensive subjective evaluation of three analysis-by-synthesis based synthesizers including (LP) synthesizer, HNM synthesizer [31] and STRAIGHT synthesizer [32], as well as three HMM-based TTS systems. The synthesized speech files were used for subjective evaluation using the recently standardized methodology for evaluating the quality of synthesized speech generated by these synthesizers based on ITU-T P.35 [26].

The subjective listening tests were designed according to ITU-T recommendation P.35 and were conducted by signal processing department in Institute for Infocomm Research.

TABLE I
DESCRIPTION OF THE NAT SCALE USED IN THE SUBJECTIVE LISTENING TESTS

	NAT scale
Rating	Description
5	Very natural, no degradation.
4	Fairly natural, little degraded.
3	Somewhat natural, somewhat degraded.
2	Fairly unnatural, fairly degraded
1	Very unnatural, very degraded

TABLE II
DESCRIPTION OF THE SIM SCALE USED IN THE SUBJECTIVE LISTENING TESTS

	SIM scale
Rating	Description
5	Similar
4	Fairly similar.
3	Somewhat similar but not intrusive.
2	Fairly unsimilar, somewhat intrusive.
1	Different, very intrusive

TABLE III
LIST OF THE TEST SENTENCES FOR SPEECH SYNTHESIZER

Author of the danger trail, Philip Steels, etc.
Not at this particular case, Tom, apologized Whittemore.
For the twentieth time that evening the two men shook hands.
Lord, but I'm glad to see you again, Phil.
God bless 'em, I hope I'll go on seeing them forever.
Will we ever forget it.
And you always want to see it in the superlative degree.
Gad, your letter came just in time.
He turned sharply, and faced Gregson across the table.
I'm playing a single hand in what looks like a losing game.

The P.35 methodology was designed to reduce the listener's uncertainty in a subjective listening test to which component(s) of an original speech, the synthesized speech, should form the basis of their ratings of overall quality. This method instructs the listener to successively attend to and rate the synthesized speech signal on:

- the speech signal alone using a five-point scale of signal distortion which measures the naturalness between synthesized speech and original speech (NAT);
- the background noise alone using a five-point scale of background intrusiveness which measures the similarity of synthesized speech (SIM);
- the overall quality using the scale of the mean opinion score (ovrl) [1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent.

The NAT and SIM scales are described in Table I. In this paper, we use the subjective ratings along the three quality scales (NAT, SIM, OVRL) to evaluate conventional and new object measures.

III. SUBJECTIVE SPEECH QUALITY MEASUREMENT

Subjective listening test were performed for evaluating synthesizers including linear predictive (LP) synthesizer, harmonic-plus-noise-model (HNM) synthesis, the Speech Transformation and Representation using Adaptive Interpolation of Weighted spectrum (STRAIGHT) synthesis, and three HMM-based TTS systems. Ten test sentences were selected from a subset of entries in the Blizzard challenge 2008. Table III lists all used test sentences.

Speech quality rating were obtained using the MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA, [39]) protocol software. In this method, a "reference stimulus" was presented, which the unprocessed male or female speech sample, while four "test stimuli" were randomly associated with the three speech synthesizers' processed stimuli, the original signal itself ("hidden reference") respectively. Participants successively used one of the three five-point rating scale (NAT, SIM, and OVRL) to listen to each of these stimuli multiple times by clicking on the corresponding icons of the software GUI and adjust the rating sliders such that a satisfactory quality rating of all stimuli was achieved.

The subjects were selected from final year project students. They have not directly involve in the work connected with assessment of the performance of speech synthesizers, or related work. They have not participated in any subjective test before they begin to work on their final year projects in our Lab. There are 6 males and 3 females aged from 19 - 24 year old. All the ten sentences were synthesized by all the tested synthesizers and recorded by female (US slt) and male (US bdl). A total of sentences was played from the PC to the test participant via Sennheiser HMD 510 closed-system headphones in office conditions and at a volume level they could adjust themselves. The tests lasted approximately 1.5 hours. Listeners took short breaks (10 minutes) between sessions. At the beginning of Session 1, the listeners were presented with a practice block of 12 trials to familiarize them with the task and the timing in the trial presentation. The practice blocks were also designed to present the listeners with the range of conditions that would be involved in the tests on both the Signal and the Back-ground scales. For each test, the panels were presented with trials in which the rating scale order of NAT, SIM, OVRL was random.

We conducted detailed statistical analysis of the subjective data collected through this procedure. A high degree of intra- and inter-rater reliability was observed with the speech quality rating by normal listeners. This database was used to benchmark the performance of different objective speech quality estimators in predicting the quality of synthetic speech.

IV. OBJECTIVE SPEECH QUALITY MEASUREMENT

Objective measures investigated in this paper belong to "intrusive" category, where features extracted from synthetic speech and its original version are compared to quantify the degree of perceptual overall difference or similarity. They can be classified into three groups: 1) metrics based on speech waveform and spectral, 2) metrics based on speech production model parameters, and 3) metrics based on the comparison of "internal representations" obtained with computational models of auditory processing. The objective speech quality measures were evaluated: segmental SNR (segSNR) [37], weighted-slope spectral distance [38], LPC-based objective measures including the log-likelihood ratio (LLR), Itakura-Saito distance measure (IS), as well as pesq measure.

A. *Time-Domain and Frequency-Weighted SNR Measures*

The time-domain segmental SNR (segSNR) measure is for computing the average signal-to-noise of processed signal [37]. The frames with segmental SNR in the range of -10 to 35 dB were considered in the average.

Weighted Spectral Slope Measure Another conventional measure is the Weighted Spectral Slope Measure (WSS) measure. The WSS is a per-frame measure in decibels and is estimated as follows [38]

$$dis_{wss}(j) = K_{spl}(K - \hat{K}) + \sum_{k=1}^{25} w_a(k)(S(k) - \hat{S}(k))^2 \quad (1)$$

where $K, \hat{K}$ are related to overall sound pressure level of the original and synthetic speech, and K_{spl} is a parameter which can be varied to increase overall performance. The WSS measure uses a psycho-acoustically motivated bank of 25 critical-band filters to estimate the smoothed short-time speech spectrum and weights the differences between the spectral slopes in each band $S(k)$ of original signal and $\hat{S}(k)$ of unprocessed signal, respectively.

B. *LPC-based Objective Measures*

The objective measures under the first group include are LPC-based objective measures including the IS, the LLR, and the cepstrum distance measures.

Itakura-Saito Distance (ISD) is defined as [34]

$$dis(\mathbf{a}_p, \mathbf{a}_o) = \frac{\sigma_p^2}{\sigma_o^2}\frac{\mathbf{a}_o R_o \mathbf{a}_o^T}{\mathbf{a}_p R_o \mathbf{a}_p^T} + \log(\frac{\sigma_o^2}{\sigma_p^2}) - 1 \quad (2)$$

where a_o is linear prediction (LP) coefficient vector of an original clean frame of speech, a_p is processed speech coefficient vector, σ_p^2 and σ_o^2 represents the gains for the processed and original clean speech respectively.

Log-Likelihood ratio measure is expressed as follows [35]

$$dis_{LLR}(\mathbf{a}_p, \mathbf{a}_o) = \log \frac{\mathbf{a}_o R_o \mathbf{a}_o^T}{\mathbf{a}_p R_o \mathbf{a}_p^T} \quad (3)$$

The LLR measure is also referred to as the Itakura distance, which measures differences in general spectral shape versus overall gain offset.

Software implementations of IS, LLR, and WSSD were taken from a toolbox provided by Hensen and Pellom [37].

C. *PESQ*

Perceptual evaluation of speech quality (PESQ; ITU-T recommendation P.862 [25]) may be the most popularly used objective measure, which incorporates a perceptual model and a cognitive modeling process for speech quality assessment of 3.2 kHz (narrow-band) handset telephony and narrow-band speech coders. The PESQ technique compute the frame-by-frame internal representation of synthetic and reference signal by first computing the power spectra, and then applying frequency and intensity warping functions based on Zwicker's loudness model [25]. Finally a cognitive model simulate listening tests, models the difference. The output is a single value in -1 to 4.5 range. As described in [25], the PESQ score is computed as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values A_{ind} as follows:

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \quad (4)$$

where $a_0 = 4.5, a_1 = -0.1$, and $a_2 = -0.0309$.

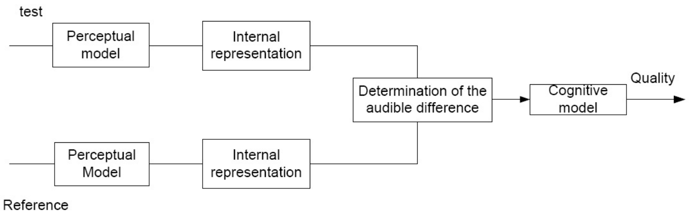

Fig. 1. Block diagram of the basic model approach for an objective perceptual measurement

Figure 1 shows the block diagram of objective measures based on the comparison of ”internal representations” obtained with computational models of auditory processing.

V. Experimental Results

Pearson's correlation coefficients will be calculated between the subjective quality measure S and the objective measure O for all tests according to

$$\rho = \frac{\sum_{i=1}^{n}(S_i - \overline{S})(O_i - \overline{O})}{[\sum_{i=1}^{n}(S_i - \overline{S})^2]^{1/2}[\sum_{i=1}^{n}(O_i - \overline{O})^2]^{1/2}} \quad (5)$$

where $\overline{S}$ and $\overline{y}$ are the mean of x and y,respectively.

The standard deviation of the error when the objective measure is used instead of subjective measure, and is given by

$$\sigma_e = \sigma_s\sqrt{1 - \rho^2} \quad (6)$$

where σ_s is the standard deviation of S, and σ_e is the computed standard deviation of the error. A smaller value of σ_e indicates that the objective measure is better at predicting subjective quality.

The Correlation coefficients and estimates of the standard deviation of the error were computed for each objective measure and each of the three subjective rating scales (NAT, SIM, OVRL). Two types of correlation analysis were performed.The first analysis was carried out by following the analysis process proposed in [27]. A total of 120 processed speech samples were used for correlation analysis including two genders (female and male) synthetic speech generated by 6 different synthesizers. The subjective ratings for each sample were averaged across all listeners involved in that test. A total of 3240 ($= 120 \times 9$ listeners $\times 3$ rating scales) subjective listening scores were used in the computation of correlation coefficients for the three rating scales. Knowing that the above correlation analysis is rather strict, we perform

TABLE IV

ESTIMATED CORRELATION COEFFICIENTS $|\rho|$ OF OBJECTIVE MEASURES

	segSNR	WSS	PESQ	LLR	IS
NAT	0.3942	0.3914	0.7825	0.1562	0.3986
SIM	0.1943	0.3748	0.3188	0.0567	0.1334
OVRL	0.3723	0.3256	0.7782	0.1460	0.3493

TABLE V

STANDARD DEVIATIONS OF THE ERROR (σ_e) FOR THE CORRELATION COEFFICIENTS $|\rho|$ OF OBJECTIVE MEASURES

	segSNR	WSS	PESQ	LLR	IS
NAT	0.7251	0.7261	0.4913	0.7793	0.7236
SIM	0.7740	0.7315	0.7478	0.7877	0.7819
OVRL	0.7323	0.7460	0.4955	0.7805	0.7393

correlation analysis using objective scores which were averaged across each condition. This analysis involved the use of mean objective scores and ratings computed across a total of 14 conditions (7 synthesizers [1] ×2 genders). Table IV and V show the correlation coefficients and standard deviations of error of the objective measures with the subjective scores of speech naturalness, speech similarity, and overall quality, respectively. From table IV, the WSS measure performed the best in terms of predicting signal similarity. In terms of speech naturalness and overall quality, the PESQ measure performed the best among these measures.

VI. DISCUSSION

Similar to P.835 process of rating the signal and background of noisy speech, we would design a process of rating the naturalness and similarity of synthetic speech to guide the listener to integrate the effects of both the signal and artifacts in making their ratings of overall quality. Of great interest of this process is finding out the contribution of individual acoustic cues in human perceptual processing of synthesized speech. According to our previous subjective data, we observed that listeners were influenced more by speech naturalness when making quality judgments. For high natural speech, listeners make overall quality judgments based on the similarity. The paper [27] confirmed our observation that the effects of different acoustic cues was different in human perceptual processing of synthetic speech, but listeners integrate these effects while making their rating. Listeners place more emphasis on the naturalness of synthetic speech itself rather than on the similarity of speech, while making the rating of overall quality.

VII. CONCLUSION

In this paper, we presented results from a study on quality prediction of synthetic speech by different objective speech and audio quality measure. The test conditions consist of synthesis speech generated by 6 synthesis algorithms for two genders (female and male). Our experimental results show that most the current objective measures are not suitable in predicting the subjective quality of synthetic speech. Further research will improve the correlations of these objective measures with subjective speech quality. Research on this topic is driven by the fact that a successful objective speech quality model will have the potential to serve as a valuable supplemental tool for the fitting and evaluation of synthesis algorithms.

[1] The original sentences were also included

REFERENCES

[1] Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt, From Text to Speech: The MITalk system. Cambridge University Press: 1987. ISBN 0-521-30641-8.

[2] http://festvox.org/blizzard/index.html

[3] D. Hirst, A. Rilliard, and V. Aubergė, Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis, in ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998.

[4] M. Vainio, J. Järvikivi, and S. Werner, Effect of prosodic naturalness on segmental acceptability in synthetic speech, in IEEE Workshop on Speech Synthesis, Santa Monica, California, 2002.

[5] P. Allen and S. Scollie, "Stimulus set effects in the similarity ratings of unfamiliar complex sounds, JASA, vol. 112, no. 1, pp. 211218, 2002.

[6] C. T. Best, B. Morrongiello, and R. Robson, Perceptual equivalence of acoustic cues in speech and non-speech perception, Percep. & Psychophys., vol. 29, no. 3, pp. 191211, 1981.

[7] L. A. Christensen and L. E. Humes, Identification of multidimensional stimuli containing speech cues and the effects of training, JASA, vol. 102, no. 4, pp. 22972310, 1997.

[8] C. Mayo and A. Turk, Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions, JASA, vol. 115, pp. 31843194, 2004.

[9] C. Mayo, R. Clark, and S. King, Multidimensional scaling of listener responses to synthetic speech, in Ninth European Conference on Speech Communication and Technology, ISCA, 2005.

[10] S. Quackenbush, T. Barnwell, and Clements, Objective measures of speech quality, Englewood Cliffs, NJ: Prentice-Hall, 1988.

[11] L. Thorpe and W. Yang, Performance of current perceptual objective speech quality measures, in Proc. IEEE Speech Coding Workshop, 1999, pp. 144146.

[12] T. H. Falk and W. Chan, Single-ended speech quality measurement using machine learning methods, IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 6, pp. 19351947, Nov. 2006.

[13] L. Malfait, J. Berger, and M. Kastner, P.563-the ITU-T standard for single-ended speech quality assessment, IEEE Trans. Audio, Speech,Lang. Process., vol. 14, no. 6, pp. 19241934, Nov. 2006.

[14] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2001, vol. 2, pp. 749752.

[15] R. Kubichek, D. Atkinson, and A. Webster, Advances in objective voice quality assessment, in Proc. Global Telecomm. Conf., 1991, vol. 3, pp. 17651770.

[16] D. Sityaev, K. Knill, and T. Burrows, Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems, in: Proc. 9th Int. Conf. on Spoken Language Process. (Interspeech 2006 ICSLP), Pittsburgh PA, 1077-1080, 2006.

[17] Viswanathan, M. and Viswanathan, M., Measuring Speech Quality for Text-to-Speech Systems: Development and Assessment of a Modified Mean Opinion Score (MOS) Scale, Computer Speech and Language 19:55-83, 2005.

[18] Cernak, M. and Rusko, M., An Evaluation of Synthetic Speech Using the PESQ Measure, in: Proc. European Congress on Acoustics, 2725-2728, 2005.

[19] Mariniak, A., A Global Framework for the Assessment of Synthetic Speech Without Subjects, in: Proc. 3rd Europ. Conf. on Speech Process. And Technology (Eurospeech93), Berlin, 1683-1686, 1993.

[20] Falk, T.H. and Mller, S., Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems, IEEE Signal Processing Letters 15: 781-784, 2008.

[21] Falk, T.H., Mller, S., Karaiskos, V. and King, S., Improving Instrumental Quality Prediction Performance for the Blizzard Challenge, in: Proc. Blizzard Challenge Workshop, Brisbane, 6 pages, 2008.

[22] Mller, S., Kim, D.-S. and Malfait, L., Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models, Acta Acustica united with Acustica 94:21-31, 2008.

[23] ITU-T Contr. COM 12-180, Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing, Source: Federal Republic of Germany (Authors: S. Mller, T.H. Falk), ITU-T SG12 Meeting, 22-29 May 2008, Geneva.

[24] Karaiskos, V., King, S., Clark, R. and Mayo, C. [2008], The Blizzard Challenge 2008. URL: http://festvox.org/blizzard/bc2008/summary Blizzard 2008.pdf

[25] ”Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of 3.1 khz Handset Telephony (Narrow-Band) Networks and Speech Codecs,” International Telecommunications Union, Geneva, Switzerland, 2001, ITU-T Recommendation P.862.

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

[27] Y. Hu and P. Loizou, Evaluation of objective measures for speech enhancement, in Proc. Interspeech, 2006, pp. 14471450.

[28] J. Tribolet, P. Noll, B. McDermott, and R. E. Crochiere, A study of complexity and quality of speech waveform coders, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 1978, pp. 586590.

[29] ”Perceptual non-instrusive single-sided speech quality measure” International Telecommunications Union, Geneva, Switzerland, 2004, ITU-T Recommendation P.563.

[30] E. Moulines, and J. Laroche, ”Non-Parametric techniques for pitch-scale and time-scale modification of speech,” Speech Commun., 16, pp. 175-205, 1995.

[31] Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. Proc. EUROSPEECH, 1995.

[32] H. Kawahara, ”Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited”, In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 1303-1306, April 1997, Munich, Germany.

[33] ”Methods for objective and subjective assessment of quality,” International Telecommunications Union, Geneva, Switzerland, 1996, ITU-T Recommendation P.800.

[34] F. Itakura and S. Saito, ”A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies,” in Speech Analysis, R. W. Schafer, Markel, J. D. (Ed.). IEEE Press. New York, 1979.

[35] F. Itakura, ”Minimum prediction residual principle applied to speech recognition,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 23(1), pp. 67 - 72, Jan. 1975

[36] S. Quackenbush, T. Barnwell, and M. Clements, Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs/New Jersey, 1998

[37] J. Hansen and B. Pellom, ”An effective quality evaluation protocol for speech enhancement algorithms,” in Proc. ICSLP ’98, Sydney, Australia, 1998

[38] D. Klatt, ”Prediction of perceived phonetic distance from critical-band spectra: A first step,” in Proc. IEEE ICASSP 1982, pp. 1278-1281

[39] ”Method for the subjective assessment of intermediate quality levels of coding systems,” International Telecommunications Union, Geneva, Switzerland, 2003, ITU-R Recommendation BS.1534.

[40] R. Huber and B. Kollmeier, ”PEMO-QłA New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception,” IEEE Trans. Audio, Speech, Lang. Process., 14(6): 1902-1911, Nov. 2006

[41] T. Dau, B. Kollmeier, and A. Kohlrausch, ”Modeling auditory processing of amplitude modulation: I. Modulation Detection and masking with narrowband carriers,” J. Acoust. Soc. Amer., vol. 102(5), pp. 2892- 2905, 1997.

[42] J.P. Marques de S$\acute{a}$, Applied Statistics Using SPSS, Statistica and matlab. New York:Springer-Verlag Berlin Heidelberg, 2003.

[43] B. Grundlehner, J. Lecocq, R. Balan, and J. Rosca, Performance assessment method for speech enhancement systems, in Proc. 1st Annu. IEEE BENELUX/DSP Valley Signal Process. Symp., 2005.