Miroslava Mrvova – Peter Pocta *

QUALITY OF SYNTHESIZED SPEECH: IMPACT OF THE NEWEST CODING APPROACHES

This contribution deals with the issue of quality of synthesized speech. It introduces principles and approaches of creating this type of speech and basic methods and techniques used to assess the quality of synthesized speech. This article also offers a short overview of relevant experimental studies discussing issues related to this kind of speech and its quality assessment. Finally, it investigates effect of the newest coding approaches (e.g. Speex, iLBC, EVRC-B, etc.) on quality of naturally-produced speech and synthesized speech (generated by diphone and unit-selection synthesizers) predicted by two different objective models and provided by subjective tests.

Keywords: *synthesized speech, synthesizer, text-to-speech systems, quality assessment, coding approaches, degradation.*

1. Introduction

In recent years, synthesized speech achieves massive increase of interest in the case of development and utilization. The reason might be the fact that speech is the most natural human form of communication and therefore there are efforts to imitate human voices. Systems used for speech synthesis offer wide range of utilization, because of their level of maturity, which allows them to be integrated for example in a place where other way of communication can not be used or in the human computer interaction systems involving higher number of modalities. Therefore the synthesized speech is implemented in many applications of daily life where this kind of speech replaces real human speaker. The synthesized speech is mainly deployed, for example, in systems providing reports containing frequently changing and routine information (weather forecast, timetable), in systems offering different dialogue situations (games) or reading various scripts (SMS-reader, e-mail reader).

In contrast to naturally-produced speech, synthesized speech represents artificially made speech, i.e. given text utterance spoken by computer. It is created by unifying pieces of speech, recorded by speaker and stored in speech database. These systems are also termed as speech synthesizers. They are based on transformation technology called text-to-speech systems (TTS). In order to realize this transformation, TTS consists of many algorithms and modules. Fig. 1 shows schematic representation of text-to-speech system.

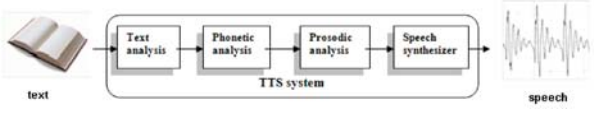

Fig. 1 Schematic representation of text-to-speech system (TTS)

In principle, functions of the TTS system can be divided into the following parts:

- Text analysis (normalization) – performs analysis of the text, which is separated into sentences. Numbers, abbreviations, symbols are replaced by their own word transcription,
- Phonetic analysis – transforms the text to voice (phonemes),
- Prosodic analysis – applies prosodic language characteristics to the selected phonemes, such as melody, speaking rate, volume, emphasis, pauses, accent, etc.
- Synthesis of the speech – generates speech signal from given sequence of prosodically-modified phonemes.

Nowadays, there are three different approaches available to create this type of speech. The currently most widely used of them is concatenate synthesis, which is restricted to speech signal and based on combining short speech strings to form a longer one. Output of this synthesis is the most naturally sounding synthesized speech. There are three main types of concatenate synthesis: unit-selection synthesis, which uses large database of recorded pieces of speech, such as words, phrases, sentences, etc. This synthesis produces voices, which are mostly indistinguishable from naturally-produced ones. Second type is diphone synthesis. The database used for this purpose consists of all diphones found in particular language. In contrast to former approach (unit-selection synthesis), overall quality of diphone synthesis is generally worse. Finally, last approach is domain-specific synthesis; its database consists of pre-recorded words and phrases, which makes it restricted to certain scripts.

* **Miroslava Mrvova, Peter Pocta**
Department of Telecommunications and Multimedia, Faculty of Electrical Engineering, University of Zilina, Slovakia, E-mail: miroslava.mrvova@fel.uniza.sk

Other approach is formant synthesis (widely deployed in past), which is based on the fundamental frequencies of amplitude spectrum of voice (formants). Systems deploying this synthesis generate artificial, robotic sounding speech (with constant quality), which cannot be confused with naturally-produced speech. Lastly, articulation synthesis represents new approach, which deals with straight human vocal track imitation, i. e. overall speech generation process. Synthesis is focused on providing isolated sounds, phones, simple words, etc. This approach has been poorly investigated at the cost of its complexity.

Definitively, synthesized speech should be indistinguishable from the human actual speech. It should characterize the most reliable copy not only in case of quality as well as in speaking style. There are efforts to ensure that synthesized speech will be the most natural, not fatiguing, not monotonous, and does not make efforts with respect to listening or comprehension [1].

For determining the output subjective quality of TTS systems (voice output devices), an application-oriented listening-only test ITU-T Recommendation P.85 [2] is recommended to be used. In general, ITU-T Recommendation P.85 is based on opinions of group of test subjects (at least 24 people), who listen to given synthesized samples and fill out the questionnaires. This recommendation defines the following rating scales: overall impression, acceptance, listening effort, comprehension problems, articulation, pronunciation, speaking rate and voice pleasantness. Assessment is based on rating called MOS (Mean Opinion Score), which represents the average values representing opinions of testing subjects or efforts needed to listen to synthesized speech expressed on the 5-point quality scale varied from bad (1) to excellent quality (5). The speaking rate uses 5-point scale varied from too slow (1) to too fast (5) and the acceptance uses only 2-point scale (yes – no). Each sample is played twice to each test subject. In first phase subjects answer questions on the information found in samples (e.g. train number, price the item). In second phase subjects are asked to assess the speech quality using one or more rating scales. For assessing the quality, two types of questionnaires, namely type I (Intelligibility) and Q (Quality) are used. Although, this method has been criticized for its shortcomings [3], [5], [25]; it is still frequently used for overall assessment of the speech output of TTS systems; but when such output is impaired by transmission degradations, a slightly modified version of this method or classical test according to ITU-T Recommendation P.800 [4] are mainly deployed.

In general, the quality of synthesized speech is evaluated in terms of intelligibility (how well the listener understands given samples) and naturalness (overall speech quality assessment). SUS (Semantically Unpredictable Sentences) belongs to the group of famous intelligibility tests. The semantically nonsense sentences with correct syntax are presented to subjects and their task is to correct the presented sentences. Each utterance is played only once. The most widespread naturalness test is MOS see details above (ITU-T Rec. P.85). Other example of naturalness test is Paired Comparison test (PC), where each sample is presented to subjects in two variants. Listener task is to choose one, which he prefers. Common to all these methods is that they are based on listener's judgments, which makes them inappropriate in terms of time and finance. Authors in [5], [6], [7] investigated the performance of the methods used for subjective assessment of quality of synthesized speech, especially the accuracy and reliability of approach defined in ITU-T Rec. P.85. In [5], the approach presented in ITU-T Rec. P.85 was compared with other available methods (test of intelligibility (SUS) and test of naturalness (MOS)) for evaluation of text-to-speech systems. Their aim was to investigate whether this approach provides the better performance than SUS and MOS test. Results showed that SUS test provides more rigorous measure of which systems were more intelligible than the other tests. However, the SUS revealed more errors which could be grouped. Overall, the ITU test is more suitable for testing intelligibility of specific application than a general purpose test. In particular, the reliability of this standard for evaluation of text-to-speech systems was investigated in [6]. Authors examined how the ranking of TTS is changing across different text genres and listening sessions. Outputs were compared with pair-comparison test (PC), using above mentioned aspects. In terms of reliability, both tests (P.85, PC) showed very similar results (from absolute score and ranking perspective). In terms of selectivity, there were minor differences between the systems across genres. In [7], the authors have compared naturally-produced speech and synthesized speech with respect to type of the speaker (male, female). Overall, female human voice was rated more persuasive and livelier than synthetized voice. Moreover, synthesized speech spoken by female speakers was rated worse in contrast to male synthesized voice. Finally, they have observed gender stereotyping effects where the results revealed that female listeners assessed male voices more favorably than vice-versa.

In order to make evaluating the perceived quality of synthesized speech more effective is necessary to have instrumental tools. Such tools should be able to predict the quality as it would be judged in an auditory tests by test subjects. At this moment, there are not available standardized models (tools) for objective quality assessment of synthesized speech. However, there are ongoing research efforts dealing with this issue, e.g. works presented in [8–10]. In order to design a new instrumental quality measure for text-to-speech systems (for both male and female synthesized speech), authors try to combine different approaches. In [8] model is based on hidden Markov models (HMM) trained on naturally-produced speech. In [9], HMM-based comparison of features extracted from synthesized signal with parametric description of the synthesized speech signal (parameters from ITU-T Rec. P.563 and parameters related to vocal expression patterns) is used in this approach. In [10], the approach presented in [9] was evaluated on auditory test databases from the Blizzard Challenges 2008 and 2009.

Oppositely, there are also ongoing efforts to verify whether the existing models designed to assess the quality of naturally-produced speech, like PESQ (Perceptual Evaluation of Speech Quality) [11–13], P.563 [14], [15], ANIQUE+ (Auditory Non-Intrusive Quality Estimation Plus) [16], [17], are capable to predict the quality of synthesized speech to a certain degree [18–20], [24], [25]. In order to realize this, many experiments were performed.

For instance, in [18], intrusive model PESQ was applied to assess the quality of synthesized speech. Authors concluded that PESQ model can be used for evaluation of synthetized speech without usage of subjective tests. On the other hand, PESQ can not be deployed for small size of diphone samples. The behavior of nonintrusive model P.563 in case of assessment of synthesized speech is investigated in [8], [19–22], [25]. Based on the results presented in [19], P.563 is better for predicting impact of transmission channel on quality of naturally-produced voice, however it has lower accuracy in prediction of the overall voice quality. Furthermore, P.563 achieves low correlation with subjective quality ratings for synthesized speech (especially in case of female synthesized voices [22]). In [20], the authors provide an explanation for this low correlation which can result from the proposed optimization of feature combinations and mapping functions in order to improve a performance of P.563 model for predicting the quality of synthesized speech. In [21] the performance of the original and modified P.563 model was also tested on synthesized speech data obtained in Blizzard Challenges 2007 and 2008. Experimental results have revealed that the algorithm, using the proposed modifications attains noticeable improvements in comparison to the original one.

Finally, there are also available studies dealing with the impact of various speech quality impairments (like noisy-type degradations, low bit rate codecs, etc.). In [23], Sebastian Moeller focused on the following issue: whether the impact of the transmission channel on the quality of synthesized speech is different from the impact on naturally-produced speech. The investigation was focused on e.g. noisy-type degradations which affected the quality of both synthesized and naturally-produced speech in the same amount; and on low bit rate codecs, which had a bit different impact on the quality of both kinds of speeches. Noisy codecs (e.g. G.726, G.728) cause more significant impact on the overall quality of synthesized speech than the artificially sounding codecs (e.g. G.729, IS-54). The signal-based comparative models, such as PESQ, TOSQA (Telecommunication Objective Speech Quality Assessment) have been applied for prediction of the quality of synthesized and naturally-produced speech impaired by low bit rate codecs. Variances in results between this models and auditory test are more considerable for synthesized than naturally-produced speech. Basically, PESQ and TOSQA are also capable to predict the quality of transmitted synthesized speech to certain degree. PESQ provides a good approximation of the quality degradation to be expected from circuit noise, whereas TOSQA model underestimates the quality at high noisy levels [24]. In [25], the authors also compared the results from various auditory tests with the predictions provided by three single-ended models (P.563, Psytechnics, ANIQUE+) using naturally-produced and synthesized voices. The samples used in this study were transmitted through different telephone channels (same impairments as used in study published in [23]). Test realized in [25] revealed that these models provide distinct correlation with results of auditory tests in the case of particular experiments.

The rest of the paper is organized as follows: Section 2 describes the investigation of impact of the newest coding approaches on speech quality in case of naturally-produced and synthesized speech usage (experimental description). In Section 3, the experimental results are presented and discussed. Finally, Section 4 concludes this paper.

2. Description of experiment

The signals transmitted through modern telephone networks are impacted by amount of degradations. Traditional, connection-based networks (analogue or digital) are affected by noise, loss, frequency distortion. Non-linear distortions from low bit-rate coding-decoding processes, talker echoes resulting from the delay, overall delay due to signal processing equipment, or time-variant degradations linked to packet or frames loss are examples of transmission degradations for new types of networks (mobiles or IP-based ones). A combination of all these impairments will be encountered when different networks are interconnected to form a transmission path from the service provider to the user. Thus, the whole path has to be taken into account for determining the overall quality of the service operated over the transmission network. As mentioned above, one of the new impairments introduced by mobile or IP-based networks is non-linear distortion from low bit-rate coding-decoding processes. Currently, this degradation is poorly investigated, especially with respect to its influence on synthesized speech [23]. This fact motivated us to investigate the impact of this distortion on speech quality. In particular, here we focus on an impact of newest coding approaches (e.g. Speex, iLBC, EVRC-B, etc.) on speech quality predictions provided by PESQ and P.563 in case of naturally-produced and synthesized speech usage.

2.1. Reference signals and experimental scenario

In this experiment, three sentences in Slovak language with length of 12 seconds were used as reference signals. Two synthesized speech signals generated with two different TTS systems (male voices) and one naturally-produced signal (recorded in an anechoic environment; with non professional male speaker) are under consideration. The decision about using male voice came from the previous study published in [7]. The tests have proved that the message produced by the male synthetic voice was rated as more favorable (e.g. good and more positive) and was more persuasive, in terms of the persuasive appeal, than the female synthetic voice. These particular differences are perceptual in nature, and more likely due to differences in synthesis quality between male and female voices.

TTS system 1 was diphone synthesizer and TTS system 2 was unit-selection synthesizer. Both systems have been developed at the Institute of Informatics of the Slovak Academy of Sciences. More about those synthesizers can be found in [26].

All speech samples have been normalized to an active speech level of −26 dB below the overload point of the digital system, when measured in accordance to ITU-T Recommendation P.56 and stored in 16-bit, 8000 Hz linear PCM; background noise was not present.

The reference signal described above were processed by following codecs ITU-T G.729AB [27] (bit rate: 8 kbps, frame size: 20 ms), ITU-T G.711 [28] (bit rate: 64 kbps, frame size: 0.125 ms), GSM-FR (GSM 06.10) [29] (13 kbps, 20ms), Internet Low Bit Rate Codec (iLBC) [30] (15.2 kbps, 20 ms), Speex [31] (4–8 kbps, 20 ms) and Enhanced Variable Rate Codec version B (EVRC-B) [32] (9.6 kbps, 20 ms). In principle the codecs used in this study can be divided into two groups. First group characterizes artificially (unnaturally) sounding codecs, such as ITU-T G.729AB, Speex, iLBC, GSM-FR and EVRC-B, whereas the ITU-T G.711 codec represents second group called naturally sounding codecs.

Speech quality was assessed by intrusive model PESQ [11-13] and nonintrusive model P.563 [14], [15]. In case of PESQ model, the raw PESQ scores were then converted to MOS-Listening Quality Objective narrow-band (MOS-LQOn) values by this equation [33]:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945*x + 4.6607}} \quad (1)$$

where x represents the obtained PESQ scores and y the computed MOS-LQOn score.

Moreover, the accuracy of PESQ's and P.563's predictions was assessed by comparing the results with subjective quality assessment.

2.2. Subjective quality assessment

As mentioned above, the obtained predictions provided by PESQ and P.563 models were compared with subjective assessments to assess their accuracy. The subjective listening tests were performed in accordance to ITU-T Recommendation P.800 [4]. Always up to 9 listeners were seated in listening chamber with reverberation time less than 190 ms and background noise well below 20 dB SPL (A). All together, 25 listeners (11 male, 14 female, age range 21-30 years, mean 24.08 years) participated in the tests. 18 of them reported to have no experience with synthesized speech. The subjects were paid for their service.

The samples were played out using high quality studio equipment in random order. Results in Opinion Score 1 to 5 were averaged to obtain MOS-Listening Quality Subjective narrowband (MOS-LQSn) values for each sample. All together, 18 speech samples were used for subjective testing of coding impact.

3. Experimental results

In this section, we present and discuss the results coming from this investigation. As mentioned above, this study focuses on a comparison of the predictions provided by objective models PESQ and P.563 with subjective scores using naturally-produced and synthesized speech, whereas different current codecs have been applied (ITU-T G.711, ITU-T G.729AB, GSM-FR, Speex, iLBC and EVRC-B) to degrade the quality of the reference signal.

Figure 2 depicts behavior of the investigated codecs on quality prediction provided by two objective models (PESQ, P.563) and by auditory tests for naturally-produced speech. We can see that artificially sounding codecs are rated significantly worse in both models' predictions compared to the auditory test. Whereas for the ITU-T G.711 codec (naturally sounding codec) the predicted quality especially provided by PESQ is in better agreement with the auditory results, as in previous case. Furthermore, P.563 model under-predicts the quality much more than PESQ in all cases.

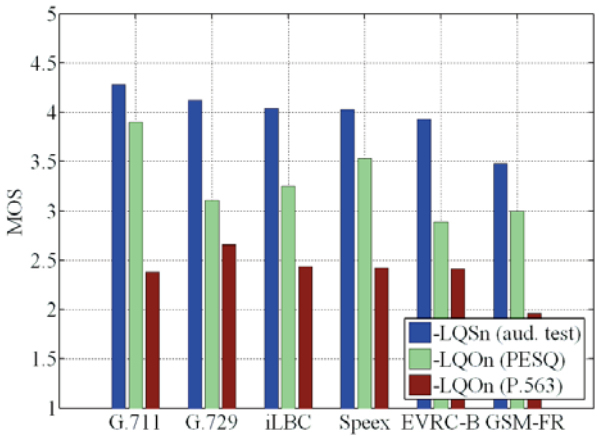

Fig. 2 Impact of the investigated codecs on MOS-LQSn and MOS-LQOn's predicted by PESQ and by P.563 in case of naturally-produced speech

Figs. 3 and 4 show the results obtained for diphone synthesizer and unit-selection synthesizer, respectively. As can be seen from Fig. 3, diphone voice (sounds less natural than unit and natural voices) was particularly disliked by test subjects. This is probably the reason for such small ratings provided by subjects. On the basis of the presented fact, we decided to omit the diphone voice from the further analysis of the behavior of synthesized speech under coding impairments. On the other hand, the behavior of the diphone voice can be used as an example how higher unnaturalness of the signal can affect the opinions of the test users. Fig. 4 depicts the effect of the investigated codecs on MOS-LQSn and MOS-LQOn predicted by PESQ as well as P.563 models for unit voice. In contrast to naturally-produced speech (see Fig. 2), the predictions of both models are in good agreement – with the exception of some predictions provided by P.563 model, like for ITU-T G.711 codec, etc. – with the auditory ratings.

Moreover, Figure 5 presents a comparison of the behavior of the synthesized speech with the behavior of naturally-produced speech from auditory ratings perspective. As can be seen from Figure 5, there are some differences between subject ratings for the synthesized speech generated by unit-selection synthesizer and naturally-produced speech. The observed differences may be due to differences in quality dimensions perceived as degradations by the test subjects. Whereas the 'artificiality' dimension introduced by the investigated 'unnatural sounding' codecs is additional degradation

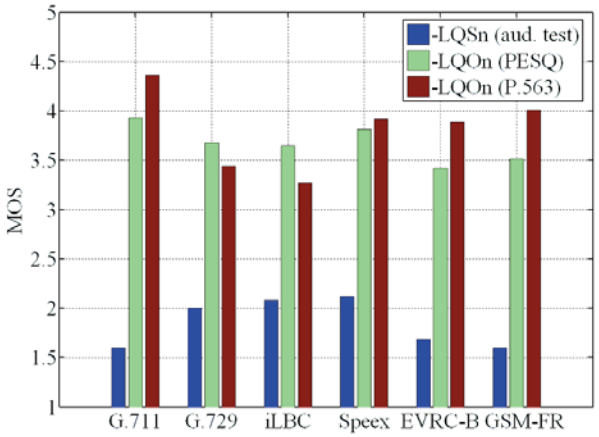

Fig. 3 Impact of the investigated codecs on MOS-LQSn and MOS-LQOn's predicted by PESQ and by P.563 in case of synthesized speech generated by diphone synthesizer

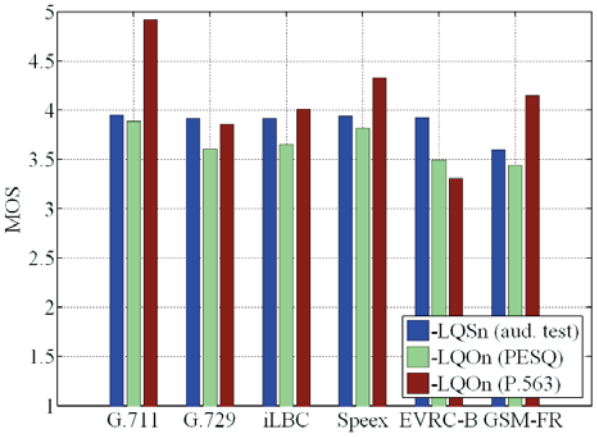

Fig. 4 Impact of the investigated codecs on MOS-LQSn and MOS-LQOn's predicted by PESQ and by P.563 in case of synthesized speech generated by Unit-selection synthesizer

for the naturally-produced speech, this is not a case for the synthesized speech, which already carries a certain degree of artificiality.

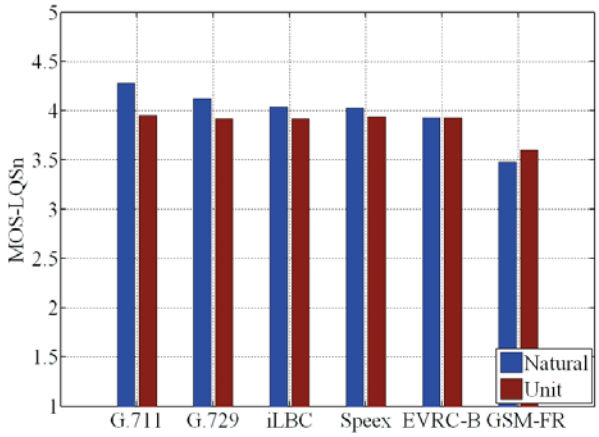

Fig. 5 Comparison between the subjective ratings for naturally-produced speech and the ratings for synthesized speech generated by unit-selection synthesizer

The results presented here are well in line with the results described in [24]. The synthesized speech is assessed a little more pessimistically than natural speech for ITU-T G.729 codec, which is shown in Figure 5.12 (p.225, [24]). On the other hand, the synthesized speech is rated a bit more optimistically by subjects than naturally-produced speech for IS-54 codec and its combinations. The effect is much more dominant for its combinations. Unfortunately, we did not investigate this codec as well as its combinations in this study but then the GSM-FR codec was involved in this study which belongs to similar family of codecs. The same behavior as for IS-54 in [24] was also reported here for GSM-FR, probably because of very similar special techniques deployed in both codec-families. Regarding the predictions of PESQ (see Figures 5.15-5.16 [24]), which were also investigated in the discussed study, they are more or less in line with our results, particularly for ITU-T G.729 codec (see Figures 2 and 4). Unfortunately, the study published in [24] is mainly focused on the different types of codecs and its combinations. This study can serve as an extension of the study published in [24].

4. Conclusion

The paper provided a brief overview of assessment of quality of synthesized speech. In addition, a overview of the current state-of-the-art of research dealing with this issue has also been described here, summarizing the experimental studies investigating the performance, accuracy and reliability of existing approaches and models (mainly designed for evaluating the quality of naturally-produced speech, but also new models designed directly for assessing the quality of synthesized speech) to evaluate the quality of synthesized speech. Finally, the paper described the experiment dealing with the impact of current codecs (ITU-T G.729AB, Speex, iLBC, GSM-FR and EVRC-B, ITU-T G.711) on the quality predicted by two objective models (intrusive PESQ, nonintrusive P.563) using naturally-produced and synthesized voices as an input signals. The obtained predictions provided by both models were compared with the ratings coming from the auditory test. The experiment revealed that the investigated codecs have a different impact on the quality of both naturally-produced and synthesized speech. Comparing the performance of both objective models, PESQ algorithm seems to be more appropriate for assessing the quality affected by the newest coding approaches than P.563 algorithm, especially in case of naturally produced speech.

Future work will focus on the following issues. Firstly, we would like to investigate the performance of a brand new ITU-T intrusive model for predicting speech quality, namely POLQA under the same conditions as investigated here (as a part of the characterization phase of this model). Secondly, on the basis of the results

obtained for the P.563 model, we have decided to try to design a new non-intrusive model for such conditions (synthesized speech and IP impairments). Thirdly, we are planning to extend the E-model towards the synthesized speech impaired by the time-varying and coding impairments.

Acknowledgement

This contribution is the result of the project implementation: Centre of excellence for systems and services of intelligent transport II., ITMS 26220120050 supported by the Research & Development Operational Programme funded by the ERDF.

"Podporujeme vyskumne aktivity na Slovensku/Projekt je spolufinancovany zo zdrojov EU"

References

[1] PSUTKA, J., MUELLER, L., MATOUSEK, J., RADOVA, V.: *We Speak with Computer in Czech (in Czech).* Academia, Praha, ISBN 80-200-1309-1, 2006, p. 752.

[2] ITU-T Recommendation P.85: *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices,* Intern. Telecommunications Union Publication, 1994.

[3] VISWANATHAN M., VISWANATHAN M.: Measuring Speech Quality for Text-to-speech Systems: Development and Assessment of Modified Mean Opinion Score (MOS) Scale. C*omputer Speech and Language 19*, 2005, p. 55–83.

[4] ITU-T Rec. P.800: *Methods for Subjective Determination of Transmission Quality,* Intern. Telecommunication Union, Geneva (Switzerland), 1996.

[5] SITYAEV, D., KNILL, K., BURROWS, T.: *Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech systems.* INTERSPEECH 2006-ICSLP, Pittsburgh, Pennsylvania, 17-21 September, 2006. Vazquez Alvarez, Y., Huckvale, M. *The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-to-Speech Systems.* In Proc. of ICSLP, 2002.

[6] VAZQUEZ, A., Y., HUCKVALE, M.: *The Reliability of the ITU-T P.85 Standard for the Evaluation of Text-to-Speech Systems.* In Proc. of ICSLP, 2002.

[7] MULLENNIX, J. W., STERN, S. E., WILSON, S. J., DYSON, C.: *Social Perception of Male and Female Computer Synthesized Speech.* Computers in Human Behavior, Vol.19, 2003, p. 407–424.

[8] FALK, T. H., MOELLER, S.: *Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech systems.* IEEE Signal Processing Letters, Vol. 15, 2008, p. 781–784.

[9] MOELLER, S., HINTERLEITNER, F., FALK, T. H., POLZEHL, T.: *Comparison of Approaches for Instrumentally Prediction the Quality of Text-to-Speech Systems.* Proc. International Conference on Spoken Language Processing (Interspeech 2010 – ICSLP), 2010.

[10] HINTERLEITNER, F., MOELLER, S., FALK, T. H., POLZEHL, T.: *Comparison of Approaches for Instrumentally Prediction the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009.* Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA), 2010, p. 1–7.

[11] ITU-T Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, International Telecommunication Union, Geneva (Switzerland), 2001.

[12] RIX, A. W., HOLLIER, M. P., HEKSTRA, A. P., BEERENDS, J. G.: *Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality, Part I - Time-delay compensation,* In J. Audio Eng. Soc., vol. 50, 2002, ISSN 1549-4950, p. 755–764.

[13] BEERENDS, J. G., HEKSTRA, A. P., RIX, A. W., HOLLIER, M. P.: *Perceptual evaluation of speech quality (PESQ) - The new ITU standard for objective measurement of perceived speech quality, Part II - Psychoacoustic model,* In J. Audio Eng. Soc., vol. 50, 2002, ISSN 1549-4950, p. 765–778.

[14] ITU-T Rec. P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications, International Telecommunication Union, Geneva (Switzerland), 2004.

[15] MALFAIT, L., BERGER, J., KASTNER, M. P.563 – *The ITU-T Standard for Single-ended Speech Quality Assessment,* In IEEE Transaction on Audio, Speech and Language Processing, vol. 14. No. 6, 2006, ISSN 1558-7916, p. 1924–1934.

[16] KIM, D.-S. ANIQUE: *An Auditory Model for Single-ended Speech Quality Estimation,* In IEEE Transaction on Speech and Audio Processing, vol. 13, No.5, 2005, ISSN 1063-6676, p. 821–831.

[17] KIM, D.-S., TARRAF, A. ANIQUE+: *A new American National Standard for Non-intrusive Estimation of Narrowband Speech Quality,* In Bell Labs Technical Journal, vol. 12, 2007, ISSN 1089-7089, p. 221–236.
[18] CERNAK, M., RUSKO, M.: *An Evaluation of Synthesized Speech Using the PESQ Measure.* Proc. Forum Acusticum, Budapest, 2005, p. 2725–2728.
[19] ITU-T Contribution COM 12 – D 174 – E. *Estimating the Quality of Transmitted Synthesized Speech with the Single-Ended Quality Prediction Model According to ITU-T Rec. P.563.* Federal Republic of Germany (Authors: S. Moeller), ITU-T SG12 Meeting, 5–13 June, Geneva, 2006.
[20] ITU-T Contribution COM 12 – C 180 – E. *Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing.* Federal Republic of Germany (Authors: S. Moeller, T.H. Falk), ITU-T SG12 Meeting, 22–29 May, Geneva, 2008.
[21] FALK, T. H., MOELLER, S., KARAISKOS, V., KING, S.: *Improving Instrumental Quality Prediction Performance for the Blizzard Challenge.* In: Proc. Blizzard Challenge Workshop, Brisbane, 2008, 6 pages.
[22] MOELLER, S., FALK, T., H. *Quality Prediction for Synthesized Speech: Comparison of Approaches.* NAG/DAGA 2009, Rotterdam, p. 1168–1171.
[23] MOELLER, S. *Telephone Transmission Impact on Synthesized Speech: Quality Assessment and Prediction.* Acta Acustica united with Acustica, Vol. 90, 2004, p. 121–136.
[24] MOELLER, S. *Quality of Telephone-based Spoken Dialogue Systems,* Springer, New York (USA), Chapter 5, ISBN 0-387-23190-0, 2005, p. 201–236.
[25] MOELLER, 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, Vol. 94, 2008, p. 21–31.
[26] DARJAA, S., RUSKO, M., TRNKA, M. *Three Generations of Speech Synthesis Systems in Slovakia,* In Proc. of XI Intern. Conference Speech and Computer (SPECOM 2006), Sankt Peterburg, 2006, ISBN 5-7452-0074-X, p. 297–302.
[27] ITU-T Rec. G.729: *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Exited Linear Prediction (CS-ACELP),* Intern. Telecommunication Union, Geneva (Switzerland), 2007.
[28] ITU-T Rec. G.711: *Pulse Code Modulation (PCM) of Voice Frequencies,* Intern. Telecommunication Union, Geneva (Switzerland), 1988.
[29] ETS 300 580-2: Digital Cellular Telecommunications System (Phase 2); Full rate speech; Part 2: Transcoding (GSM 06.10 version 4.2.1), European Telecommunications Standards Institute, 2000.
[30] IETF RFC 3951: *Internet Low Bit Rate Codec (iLBC),* Internet Engineering Task Force, 2004.
[31] VALIN, J.-M. *Speex: A Free Codec for Free Speech,* In Proc. of Australian National Linux Conference (LCA 2006), Dunedin : New Zealand, 2006.
[32] 3GPP2 C.S0014-C: Enhanced Variable Rate Codec, Speech Service Options 3, 68, and 70 for Wideband Spread Spectrum Digital Systems, Third Generation Partnership Project 2, 2007.
[33] ITU-T Rec. P.862.1: *Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO,* Intern. Telecommunication Union, Geneva (Switzerland), 2003.