No. 2 · pp. 181 – 360 · March/April 2017 · E 21 466 · ISSN 1610-1928 Volume 103

ACTA ACUSTICA UNITED WITH ACUSTICA

The Journal of the European Acoustics Association (EAA) · International Journal on Acoustics

Table of Contents

Fast Track Articles

Scientific Papers

General Linear Acoustics

Ultrasonics

Structural Acoustics

Environmental Noise

Building Acoustics

Audio Signal Processing and Transducers

S. Hirzel Verlag · Stuttgart

European Acoustics Association
EAA

No. 2 · pp. 181 – 360 · March/April 2017 · E 21 466 · ISSN 1610-1928 Volume 103

ACTA ACUSTICA UNITED WITH ACUSTICA

The Journal of the European Acoustics Association (EAA) · International Journal on Acoustics

Computational and Numerical Acoustics

Hearing, Audiology and Psychoacoustics

Speech

Musical Acoustics

Virtual Acoustics

Technical and Applied Papers

Short Communications

S. Hirzel Verlag · Stuttgart

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 103 (2017) 311 – 316

DOI 10.3813/AAA.919059

Subjective and Objective Measurement of the Intelligibility of Synthesized Speech Impaired by the Very Low Bit Rate Stanag 4591 Codec Including Packet Loss

Peter Počta[1], John G. Beerends[2]
[1] Department of Multimedia and Information-Communication Technology, FEE, University of Žilina,
01026 Žilina, Slovakia. pocta@fel.uniza.sk
[2] TNO, P. O. Box 96800, 2509 JE The Hague, The Netherlands

Summary
This paper deals with the intelligibility of speech coded by the STANAG 4591 standard codec, including packet loss, using synthesized speech input. Both subjective and objective assessments are used. It is shown that this codec significantly degrades intelligibility when compared to a standard narrowband filtered version of the synthesized speech. Packet loss impact is strongly dependent on the exact timing location. Furthermore it is shown that POLQA Intelligibility, a speech intelligibility prediction model, is capable of providing good intelligibility predictions for all investigated conditions.

PACS no. 43.71.Gv, 43.72.Kb

1. Introduction

In recent years, synthesized speech has reached a level of quality which allows it to be integrated into many real-life applications, e.g. e-mail and SMS readers, etc. In particular, Text-to-Speech (TTS) can fruitfully be used in systems enabling interaction with an information database or a transaction server, e.g. via the telecommunication network [1].

Modern telecommunication networks, however, introduce a number of degradations which have to be taken into account when services are planned and developed. The type of degradation depends on the specific network under consideration. In traditional, connection-based (analogue or digital) networks, loss, frequency distortion, talker echo and noise are the most significant degradations. In contrast, new types of networks (e.g. mobiles or IP-based ones) introduce impairments which are perceptively different from the traditional ones. Examples are non-linear distortions from low bit-rate coding-decoding processes (codecs), overall delay due to signal processing equipment, talker echoes resulting from the delay in conjunction with acoustic reflections, or time-variant degradations when packets or frames get lost on the digital channel. It is worth noting that the handset nowadays introduces the most prominent portion of degradations and that currently wideband-speech coding is introduced in the mobile networks and handsets. In this study a wideband reference speech signal is coded with a narrowband codec and intelligibility is assessed for several degradations amongst which is also a telephone band (300-3400Hz) filtered version of the signal. This allows to compare the impact of band limiting and coding separately.

To quantify the intelligibility of a speech-transmission chain, a number of measurement techniques have been developed during the past decades. In the subjective domain, examples are the consonant-vowel-consonant (CVC) test [2], using three-letter nonsense words in silence, and the speech reception threshold (SRT) test [3], using short everyday sentences in noise in an adaptive procedure. Further tests are the modified rhyme test (MRT) [4] and diagnostic rhyme test (DRT) [5]. The DRT and MRT tests are typical examples of so called closed-response tests. In these tests subjects are offered a set of alternatives from which a selection has to be made. On the other hand, open response tests allow listeners to respond to what they think to have heard.

In the objective domain, the articulation index (AI) [6] and the standardized speech-transmission index (STI) [7, 8, 9, 10, 11] are worldwide adopted methods for predicting the speech intelligibility for virtually any electroacoustic situation. The STI method is a quick objective method for assessing the speech-transmission quality of transmission channels. Using the STI method, the speech-transmission index can be both measured and calculated. The STI, a value between 0 and 1, indicates how

Received 15 September 2015,
accepted 2 February 2017.

well speech is transmitted through the transmission channel with respect to intelligibility. Using the STI value, the speech intelligibility for different types of speech material (numbers, CVC words, sentences) can be predicted, deploying a customized transformation for each type of speech material. The STI method makes use of a test signal that contains spectrotemporal characteristics similar to those found in natural speech. By comparing the intensity fluctuation patterns (envelope spectra) for both the degraded output and the reference input signals, the modulation transfer function (MTF) is derived. The MTF forms the basis for quantifying how well speech information is transmitted by the transmission channel. Based on the MTF, the STI is calculated. It should be noted here that STI is not applicable in modern digital speech-transmission channels, such as those used for VoIP, due to non-linear speech processing introduced by advanced speech coding deployed in such channels, see more details in [11]. Therefore new methods for predicting the speech intelligibility in such conditions are being developed [12, 13] of which PESQ Intelligibility [12] is best suited for assessing codec degradations including packet loss. An upgraded version of the PESQ Intelligibility method, called POLQA Intelligibility, is currently being developed by Q9 of ITU-T SG12 under the work item P.OSI (series P recommendations Objective Speech Intelligibility). In fact, both PESQ Intelligibility and POLQA Intelligibility are adapted versions of speech quality prediction models, namely PESQ [14, 15, 16] and POLQA [17, 18, 19].

Some work has been carried out to study the performance of the PESQ, PESQ Intelligibility and POLQA models in predicting speech intelligibility for natural speech degraded by low bit rate codecs and packet loss. In [20], Beerends *et al.* investigated to what extend PESQ (ITU-T P.862) can be used to predict speech intelligibility with vocoders using the NATO speech intelligibility test on vocoders/noise suppressors. This database consisted of long speech files (about 3 minutes) containing 50 CVC words embedded in a carrier sentence. Twelve different noise conditions were used to assess the quality of 9 vocoders/noise suppressors. Bit rates of the codecs were between 1 and 5 kbit/s. The subjective intelligibility scores were acquired through an open response test. The results show that PESQ provides acceptable results when used to predict the speech intelligibility. It has been also shown that some modifications of the PESQ model can increase the correlation between objective and subjective intelligibility scores from 0.86 up to 0.95. Finally, the authors have concluded that further validations are necessary in order to see if the improvements that are implemented can cope with a wide range of distortions.

In [12], Beerends *et al.* further exploited the idea of using a PESQ-like modeling approach in predicting subjectively obtained intelligibility scores. They focused on a large series of degradations covering band filtering, peak clipping, reverberation, noise, analog radio distortions, low bit-rate speech coding, bandwidth limitation, different types of background noise (white, babble, car), multiplicative noise and room response distortions. The results show that it is possible to develop an objective speech intelligibility measurement algorithm on the basis of PESQ despite the fact that PESQ itself shows low correlations (around 0.5) between its raw output and the subjectively obtained intelligibility scores coming from an open response CVC test. A simple retraining of PESQ already provides a significant improvement with a correlation of around 0.8 on untrained data. By adding advanced features the correlation between objective and subjective measurements is improved significantly and the correlation on data not involved in training process is around 0.9, a level that allows practical use.

In [21], Ullmann at el. proposed a no-reference objective intelligibility-assessment approach based on comparison of phoneme posterior probability sequences. Firstly, they compared a performance of the proposed approach with speech quality predictions provided by the POLQA model using natural speech samples coming from ITU-T P.501 [22] degraded by AMR, EVRC, MELP and codec2 codecs and simulated frame losses ranging from 5 to 40%. Secondly, they evaluated the proposed approach on the 2011 Blizzard Challenge data [23], which comprises speech recordings synthesized with 12 different TTS systems. Particularly, they deployed a subset of 26 semantically unpredictable sentences [24] in English, for which subjective intelligibility scores are provided in the form of Word Error Rates (WER). They have shown that the proposed approach yields realistic results for low bit rate codec distortions, and that it is also able to assess speech intelligibility of TTS systems.

To the best of our knowledge, there is no work dealing specifically with synthesized speech impaired by very low bit rate codecs and packet loss. Conditions involving very low bit rate codecs are especially relevant for service providers and network operators dealing with bandwidth shortages, and are of very high importance for emergency-response environments. We have chosen the STANAG 4591 [25, 26] codec for this study, as this codec is involved in emergency-response environments deploying the Terrestrial Trunked Radio (TETRA) transmission system [27] and other professional mobile radio systems. We study the impact of the codec including packet loss on the intelligibility of synthesized speech experienced by a user in a closed-response CVC test. A closed-response test was chosen because accidental bias effects, e.g. caused by a certain codec degradation or packet loss condition, will be averaged out when assessing the condition score. If one degraded CVC word resembles another, leading to a low correct identification for this single word, any objective measurement method would fail to predict this single identification error. However in a closed set the average over all CVC alternatives would average out this bias effect, allowing to predict the condition score with a perceptual measurement approach. Open response tests always suffer from this, unpredictable, language dependent, bias effects, see a study published in [28] for more detail. The synthesized speech samples are generated with two different types of TTS systems, namely with a unit-selection synthesizer and an HMM synthesizer. Moreover, we also check the performance of the current version of the POLQA Intelligibility model for the investigated conditions. The performance of the POLQA Intelligibility model is assessed by comparing the predictions with subjective intelligibility scores obtained from the test described in this paper. The aim of this study is two-fold: firstly, we would like to know to what extent degradations introduced by the STANAG 4591 codec and packet loss have an impact on the intelligibility of synthesized speech. Secondly, we would like to see whether the POLQA Intelligibility model is able to provide valid predictions of perceived intelligibility for the given application domain.

The remaining of the paper is organized as follows. Section 2 describes the subjective test carried out within this study and its results. In Section 3, the experimental results obtained from the prediction model are compared with the subjective data presented in this paper and discussed. Section 4 provides the final conclusions.

2. Subjective test

The first aim of this study was to quantify the impact of degradations introduced by the STANAG 4591 codec, including packet loss, on the intelligibility of synthesized speech. To this end, a subjective test has been carried out. The following sections provide a description of this test and the results that are obtained.

2.1. Experiment description

In this study, a closed-response test was carried out with short nonsense three-letter CVC words. In all experiments, up to 2 listeners were seated in a small listening room (acoustically treated) with a background noise below 20 dB SPL (A). All subjects were Slovak Nationals whose first language was Slovak. The subjects were remunerated for their efforts. The speech samples were played out using high-quality studio equipment in a random order and diotically presented over Sennheiser HD 455 headphones (presentation level: 73 dB SPL (A)) to the test subjects. The subjects listened to the stimuli and choose an alternative from a list of alternatives. For each correct assessment, a score of one was given and for an incorrect or skipped/not recognized assessment, a score of zero was given. The average score over all subjects represent the subjective intelligibility score for a particular degraded word. The value of a subjective score lies between 0% (not intelligible) and 100% correct (perfectly intelligible). The average score over all words represents the intelligibility of a condition.

The speech samples used in this study were generated by two state-of-the-art Slovak TTS systems (male voices) developed by the Institute of Informatics of the Slovak Academy of Sciences. The first one was a unit-selection synthesizer (Kempelen 2.1 [29], marked as ‘Unit’ in this paper), and the second one was a Hidden Markov Model synthesizer (Kempelen 3.0 [30], marked as ‘HMM’ in this paper). The decision to use a male voice was influenced by a previous studies published in [31, 32]. The tests published in [31] showed that there is a difference between the intelligibility of female and male talkers. More specifically, the female talkers were generally more intelligible than male talkers while in [32] it was shown that male synthetic voices were rated more favorable (e.g. good and more positive) and more persuasive, in terms of the persuasive appeal, than the female synthetic voices. These particular differences are perceptual in nature, and are most likely due to differences in synthesis quality between male and female voices.

Table I. Description of the test conditions used in the closed-response test.

No.	Description of test condition
1	Clean reference signal
2	Bandwidth limitation 300–3400 Hz (NB)
3	STANAG 4591 2.4 kbit/s [25, 26]
4	STANAG 4591 2.4 kbit/s +10% Packet loss
5	STANAG 4591 2.4 kbit/s +15% Packet loss
6	STANAG 4591 2.4 kbit/s +20% Packet loss
7	STANAG 4591 1.2 kbit/s [25, 26]
8	STANAG 4591 1.2 kbit/s + 15% Packet loss
9	Pink noise SNR 6 dB
10	Pink noise SNR 0 dB

In order to have a high discriminative power, the CVC words were chosen to sound as similar as possible, like used in MRT and DRT testing [4, 5]. The CVC words used in the test were chosen on the basis of the Chomsky and Halle feature distance matrix for English consonants; see more details in [33]. The final set consisted of the following 7 CVC words: Bit, Dit, Pit, Tit, Vit, Zit (as pronounced in the English zero) and THit (as in that). The fact that only 7 words are used in a closed set converts the word task towards a simple sound classification task making the test largely language independent. A test involving the CVC phonemes covering stops (b, d, p) and fricatives (t, v, z, th), further validations on nasals, glides and affricates (e.g. mit, wit, chit) will be carried out in a follow-up study.

Ten test conditions representing typical degradations commonly seen in very low bit rate voice communication over telecommunication networks or in emergency-response environments were investigated in this test, see Table I. 4 test conditions, namely test conditions No.1, 2, 9 and 10 (see Table I) are identical to those of the other tests ran within the P.OSI work item and allow comparison of the results obtained in this tests with the other tests. The test conditions No.1, 9 and 10 represent wideband-speech samples sampled at 22.05 kHz (the highest sampling rate offered by the TTS systems deployed in this study).

Regarding the packet loss, seven different loss locations were simulated in the speech samples in order to take an effect of loss location on the intelligibility into account. No packet loss concealment technique was deployed. Packet loss was generated by deleting short segments of speech, corresponding to an actual length of the speech codec frame. Due to the short length of the speech samples, single CVC words of length of approx. 450 ms, it was not possible to simulate a real packet loss. Therefore, it was also not possible to deploy packet loss concealment technique in the experiment, despite the fact that it is widely used in telecommunications. The same loss locations and lengths were simulated in test conditions No.5 and 8 to compare the perceived intelligibility of the STANAG 4591 codec operating at 2.4 kbit/s and 1.2 kbit/s under identical packet-loss conditions.

As it is widely known, the TTS systems can have a problem to correctly synthesize words for which they are not optimized such as nonsense CVC words. It was checked whether all the synthesized CVC words are correctly identified by 4 listeners (not involved in a main test) in a pretest. In total, 70 speech samples per TTS system (7 CVC words per test condition * 10 test conditions) were used in this test. These 70 speech samples were presented to the test subjects five times in a different random order to be in line with a testing procedure used for similar tests ran at TNO and within the P.OSI work item. The subjects that performed best in the previous tests published in [28] were invited to take part in this test. Altogether, 4 listeners (2 male, 2 female, 23-47 years, mean 31.25 years) participated in the test. As 4 test subjects were involved in the test and the samples were presented 5 times in a different random order, 20 intelligibility scores per sample were obtained in this experiment.

A separate test was run for each TTS system. Each test consisted of a three-phase training session and a test session. As THit is a very atypical word for the Slovak language, special attention was paid to properly train test subjects for this special word avoiding potential confusion in the other phases of the training session and in the test. For this first-phase training, a non-degraded version sampled at 22.05 kHz of that word was presented to the subjects. In the second-phase training, subjects had to identify two non-degraded synthesized speech samples by choosing the correct word from the list of alternatives. If they correctly identified both words, they were allowed to take part in the main test. The aim of the third-phase of the training session was to familiarize the subjects with the software tool deployed in the main test. This was carried out by simulating a short part of the main test.

2.2. Experimental results

Figure 1 presents the results of the closed-response test averaged over 140 intelligibility scores (7 words per condition * 20 votes per word). The average estimated 95% confidence interval for the true population proportion was ±8.1% for the HMM synthesizer and ±7.6% for the unit-selection synthesizer. In principle, the test conditions can be split into 4 groups, namely bandwidth limitation (No.2), STANAG 4591 codec operating at 2.4 kbit/s (No.3-6), STANAG 4591 codec operating at 1.2 kbit/s (No.7-8) and pink noise (No.9-10). Regarding an inter-group behavior, it can be clearly seen from Figure 1 that the intelligibility monotonically decreases as impairments introduced by bandwidth limitation, STANAG 4591 codec or pink noise are introduced. Interestingly, the same average intelligibility scores were obtained for the conditions No.3 and 7 (STANAG 4591 codec operating at 2.4 kbit/s and STANAG 4591 codec operating at 1.2 kbit/s) for the HMM synthesizer. Roughly the same effect was obtained for the unit-selection TTS system where a small difference of 2.1% was found. Looking at the intra-group behavior we see that the intelligibility decreases monotonically for the HMM synthesizer with increasing packet loss. For the Unit synthesizer, the test condition carrying the 20% packet loss has a better intelligibility than the one carrying only 15% loss. This is due to the bursty characteristic of the loss combined with the synthesizer characteristics, which can accidentally lead to a more severe degradation than expected.

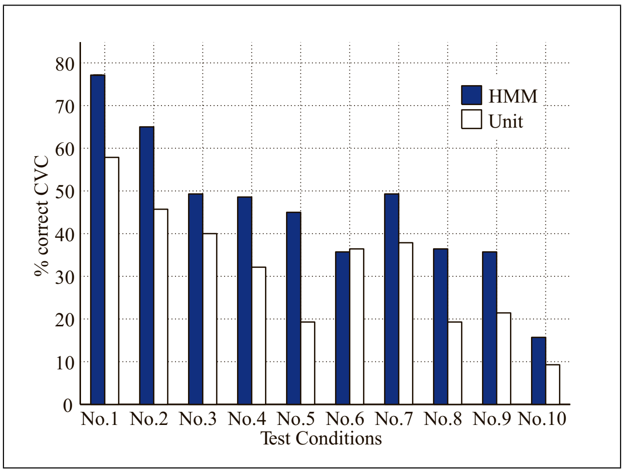

Figure 1. Effect of test conditions (see Table I for more information about the investigated test conditions) on average intelligibility scores in the closed-response test.

When comparing the STANAG 4591 codec operating at 2.4 kbit/s and 1.2 kbit/s under packet-loss conditions (test conditions No.5 and 8), we see from Figure 1 that the second one is more impaired by the packet loss even though the average intelligibility achieved for error-free conditions (test conditions No.3 and 7) is equivalent. It is also worth noting that the average intelligibility reported for the unit-selection synthesizer is the same for both investigated bit rates.

When the investigated TTS systems are compared over all the test conditions from the intelligibility perspective, we can conclude that the HMM TTS system is more robust against degradations of similar sounding CVC words than the unit-selection synthesizer.

3. Objective test

In this section, the subjective results are compared to the predictions made with the POLQA Intelligibility model. The version that is evaluated in this paper uses the basic ideas as given in [12] but then applied to POLQA [17, 18]. The comparison is performed for all the experimental conditions.

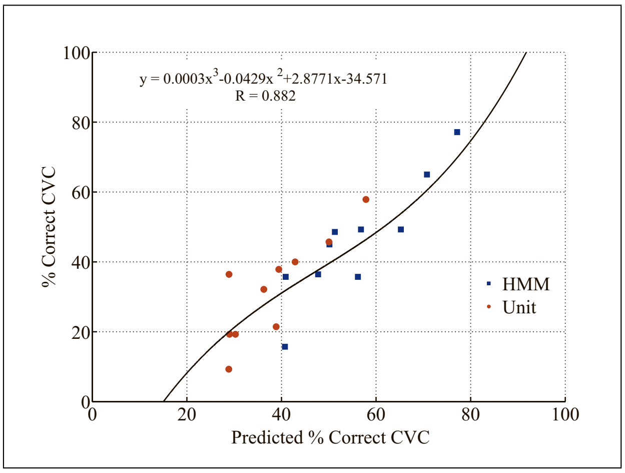

Figure 2. Correlation between the subjective results of closed-response test and POLQA Intelligibility predictions.

As the structure of the speech material used in the subjective test does not comply with the structure used in the POLQA standard, it is necessary to restructure the speech material. The speech files used in the objective measurements consist of a concatenation of the 7 individual CVC words ('Bit+Dit+Pit+Tit+Vit+Zit+THit') for both the reference and degraded files used in the test.

Figure 2 compares the subjective intelligibility scores obtained from the closed-response test described in Section 2 with the POLQA Intelligibility predictions. In order to model the experimental context of a particular experiment, 3rd order monotonic regression was used in the calculation of the correlation between the measured and predicted CVC scores. A similar approach was also used in the standardization process of the PESQ and POLQA model and allows to ignore bias and context effects which are caused by experimental context. More detail about the context-dependent mapping can be found in [34]. The correlation between the objective and subjective measurements is 0.88, a level that allows to make reliable intelligibility predictions (see Figure 2).

Moreover, it can be observed from Figure 2, the POLQA Intelligibility model provides a better differentiation of the test conditions than the subjective test.

It is interesting to note here that the intelligibility inconsistency reported for the unit-selection synthesizer in the subjective test, see Section 2 for more detail, was not reproduced by the POLQA Intelligibility model. On the other hand, this effect has occurred for the HMM synthesizer when it comes to the POLQA Intelligibility predictions but to a much lower extent as reported in the previous case. Current investigations are focused on improving this behavior of the POLQA Intelligibility model.

Table II provides an overview of all correlations between the intelligibility scores obtained from the closed-response test described in Section 2 and the intelligibility predictions obtained with other available natural and synthesized speech databases used previously in the training and validation of PESQ/POLQA Intelligibility. Detailed descriptions of the databases are available in [12, 28, 35].

Table II. Pearson correlation coefficients between the subjective scores obtained from the closed-response test and other available natural and synthesized speech databases and the POLQA Intelligibility model.

Database (speech type)	POLQA
TNO CVC standard (Natural)	0.88
Telecom distortions (Natural)	0.88
Analogue radio distortions (Natural)	0.95
Telecom distortions Closed (Synthesized)	0.94
Very Low Bit Rate Telecom Distortions Closed (Synthesized)	0.88
Average	0.91

The results presented in Table II show that POLQA Intelligibility provides a good correlation for each data even for conditions involving synthesized speech degraded by impairments introduced by very low bit rate telecommunication conditions. The average correlation over all the databases is 0.91, a level that allows practical use.

4. Conclusions

The intelligibility of synthesized speech coded by the STANAG 4591 standard narrowband codec, operating at 1.2 and 2.4 kbit/s, is significantly degraded when compared to a standard narrowband filtered version of the synthesized speech. Packet loss impact is strongly dependent on the exact timing location. POLQA Intelligibility, a speech intelligibility prediction model based on the POLQA ITU-T P.863 standard, is capable to provide valid predictions of the intelligibility for all degradations that are tested.

References

[1] S. Moeller: Telephone transmission impact on synthesized speech: Quality assessment and prediction. Acta Acustica united with Acustica **90** (2004) 121–136.

[2] H. J. M. Steeneken: Subjective phoneme, word and sentence intelligibility measures. in: On measuring and predicting speech intelligibility, pp.37-75. Ph.D. thesis, University of Amsterdam, The Netherlands, 1992, ISBN 90-6743-209-1.

[3] R. Plomp, A. M. Mimpen: Improving the reliability of testing the speech reception threshold for sentences. Audiology **8** (1979) 43–52.

[4] G. Fairbanks: Test of phonetic differentiation: The rhyme test. Journal of Acoustical Society of America **30** (1958) 596–600.

[5] W. Voiers: Diagnostic acceptability measure for speech communication systems. Proc. IEEE ICASSP, Hartford, CT, USA, 1977, 204–207.

[6] N. R. French, J. C. Steinberg: Factors governing the intelligibility of speech sounds. Journal of Acoustical Society of America **19** (1947) 90–118.

[7] H. J. M. Steeneken, T. Houtgast: A physical method for measuring speech-transmission quality. Journal of Acoustical Society of America **67** (1980) 318–326.

[8] H. J. M. Steeneken: On measuring and predicting speech intelligibility. Ph.D. thesis, University of Amsterdam, The Netherlands, 1992.

[9] ANSI S3.5: Methods for calculation of the speech intelligibility index. American National Standards Institute, New York, USA, 1997.

[10] ISO 9921: Assessment of speech communication. International Standards Organization, Geneva, Switzerland, 2003.

[11] IEC 60268-16: Sound system equipment. Part 16: Objective rating of speech intelligibility by speech transmission index. International Electrotechnical Commission, Geneva, Switzerland, 2003.

[12] J. G. Beerends, R. van Buuren, J. Van Vugt, J. Verhave: Objective speech intelligibility measurement on the basis of natural speech in combination with perceptual modeling. Journal of Audio Engineering Society **57** (2009).

[13] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen: An algorithm for intelligibility prediction of time-frequency weighted noisy speech," audio, speech, and language processing. IEEE Transactions **19** (2011) 2125–2136.

[14] A. W. Rix, M. P. Hollier, A. P. Hekstra, J. G. Beerends: Perceptual evaluation of speech quality (PESQ). The new ITU standard for objective measurement of perceived speech quality, part i - time-delay compensation. Journal of Audio Engineering Society **50** (2002) 755–764.

[15] J. G. Beerends, A. P. Hekstra, A. W. Rix, M. P. Hollier: Perceptual evaluation of speech quality (PESQ). The new ITU standard for objective measurement of perceived speech quality, part ii - psychoacoustic model. Journal of Audio Engineering Society **50** (2002) 765–778.

[16] ITU: 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, ITU-T Rec. P.862, 2001.

[17] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, M. Keyhl: Perceptual objective listening quality assessment (POLQA). The third generation ITU-T standard for end-to-end speech quality measurement. Part I: Temporal alignment. Journal of Audio Engineering Society **61** (2013) 366–384.

[18] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, M. Keyhl: Perceptual objective listening quality assessment (POLQA). The third generation ITU-T standard for end-to-end speech quality measurement. Part II: Perceptual model. Journal of Audio Engineering Society **61** (2013) 385–402.

[19] ITU: Perceptual objective listening quality assessment. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.863, 2011.

[20] J. G. Beerends, S. van Wijngaarden, R. Van Buuren: Extension of ITU-T recommendation p.862 PESQ towards measuring speech intelligibility with vocoders. Proc. New Directions for Improving Audio Effectiveness, pp. 10-1-10-6, Neuilly-sur-Seine, France, 2005.

[21] R. Ullmann, M. Magimai-Doss, H. Bourlard: Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences. Idiap research report 16-2014, October 2014.

[22] ITU: Test signals for use in telephonometry. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.501, 2012.

[23] S. King, V. Karaiskos: The blizzard challenge 2011. Proc. Blizzard Challenge Workshop, 2011.

[24] C. Benoît, M. Grice, V. Hazan: The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Communication **18** (1996) 381–392.

[25] NSA: The 600, 1200 and 2400 bit/s NATO interoperable narrow band voice coder. NATO Standardization Agency, Brussels, Belgium, STANAG No.4591, 2005.

[26] T. Wang, K. Koishida, V. Cuperman, A. Gersho, J. S. Collura: A 1200/2400 bps coding suite based on MELP. Proc. IEEE Speech Coding Workshop, 2002, 90–92.

[27] ETSI: Terrestrial trunked radio (TETRA); voice plus data (V+D); Part 1: General network design. European Telecommunications Standards Institute, ETSI EN 300 392-1, 2009.

[28] P. Počta, J. G. Beerends: Subjective and objective measurement of synthesized speech intelligibility in modern telephone conditions. Speech Communication **71** (2015) 1–9.

[29] S. Darjaa, M. Rusko, M. Trnka: Three generations of speech synthesis systems in Slovakia. Proc. XI International Conference Speech and Computer (SPECOM 2006), Sankt Petersburg, Russia, 2006, 297–302.

[30] S. Darjaa, M. Trnka, M. Cerňak, M. Rusko, R. Sabo, L. Hluchý: HMM speech synthesizer in Slovak. Proc. GCCP 2011: 7th International Workshop on Grid Computing for Complex Problems, Bratislava, Slovakia, 2011, 212–221.

[31] A. R. Bradlow, G. M. Torretta, D. B. Pisoni: Intelligibility of normal speech. I: Global and fine-grained acoustic-phonetic talker characteristics. Speech Communication **20** (1996) 255–272.

[32] J. W. Mullennix, S. E. Stern, S. J. Wilson, C.-I. Dyson: Social perception of male and female computer synthesized speech. Computers in Human Behavior **19** (2003) 407–424.

[33] N. Chomsky, M. Halle: The sound pattern of English. Harper & Row, New York, USA, 1968.

[34] ITU: Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. International Telecommunication Union, Geneva, Switzerland, ITU-T Rec. P.1401, 2012.

[35] ITU, P. Počta, J. G. Beerends: Subjective and objective measurement of synthesized speech in modern telephone conditions. ITU-T SG 12 Contribution 93, 2013.