ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2010, Vol. 39, No. 1

LITHUANIAN SPEECH RECORDS DATABASE FOR VOICE CODECS QUALITY ASSESSMENT

Šarūnas Paulikas, Mohamad Itani

Department of Telecommunication Enginering, Vilnius Gediminas Technical University
Naugarduko 41, LT03227, Vilnius, Lithuania
e-mail: sarunas.paulikas@el.vgtu.lt

Abstract. The English language has had a dominating influence on the advance of telecommunications. With many of the major developments coming from primarily English speaking areas there is the risk that these advances may not be linguistically robust. It is noted that the quality of a speech produced by voice codecs mainly is assessed using samples of English language. However, some other investigations show that influence of language on codecs performance could be noticed. This paper presents the Lithuanian speech records database that conforms to the requirements of ITU-T Rec. P.50 and is dedicated for assessment of voice codecsquality. Experimentallywe estimate thequality ofAMR andSpeex voice codecs using apresented database of Lithuanian speech records as well as American and British English and Arabic speech records databases supplied in the Appendix I of ITU-T Rec. P.50.

Keywords: speech signalprocessing, speech coding, multi-lingual speechprocessing,quality assessment.

1. Introduction

Human speeches (sounds) differ from one language to another. For example, with comparison to the English language, the Lithuanian language uses many vowels. The English language has had a dominating influence on the advance of telecommunications. With many of the major developments coming from primarily English speaking areas there is the risk that these advances may not be linguistically robust. Research carried out in [8] shows that coders interact with individual voices so that speech is degraded deferentially for different talkers. An extensive study that was carried out in [9] on English, Japanese, Finish and German languages also shows domination of English language in codecs' quality tests. Parray, Burnett and Chicharoa [11] presented a set of recommendations for codebook design for multi-lingual environments.

Now in telecommunication sector a lot of companies that design voice communication products and regulatory institutions admit importance of language on performance of voice communication systems. For example, over 20 languages (mostly from European countries) that are listed in ITU-T Rec. P.50 Appendix I were extensively used in tests of AMR codec that were chosen by 3GPP as default voice codec in 3G mobile communications [2].

Unfortunately, Lithuanian language has left aside of this investigation. One of the reasons for not including Lithuanian language in codecs performance and other tests is lack of appropriate Lithuanian speech records database [10]. It should be noticed that in Lithuania there is a few Lithuanian speech records databases created by Institute of Mathematics and Informatics, Lithuanian Radio News (LRN0 and LRN0.1), Vytautas Magnus University (VDU-RTG). Also, Kaunas University of Technology together with University of Vilnius created LITGIS database (records of Lithuanian digits sequences). However, these speech records databases are intended to use for Lithuanian speech recognition but not for testing of voice communication systems.

In this paper we present a Lithuanian speech records database that satisfies requirements of ITU-T Rec. P.50 Appendix I and is dedicated for tests of the performance of voice telecommunication systems. Also, using created database, we investigate the performance in the sense of quality of decoded speech signal of two the most popular speech codecs (Speex and AMR) [14, 13]. The quality of transformed speech signals will be estimated

using auditory listening tests (ITU-T Rec. P.830 [4]) and objective quality estimation technique PESQ (ITU-T Rec. P.862 [1]).

2. Assessment of Speech Quality

ITU-T Rec. P.830 describes methods and procedures for conducting subjective performance evaluations of digital speech codecs. Subjective testing is the most widely used method of assessing the performance of digital codecs. As a matter of fact, listening-only tests are the only feasible method of subjective testing when the transmission path is digital and/or non-linear, because of simple objective measurements are insufficient to ensure adequate transmission performance. The aim of a subjective testing methodology is to measure the degradation contributed by the non-linear part of the transmission path, and hence to ensure that the performance of the complete system is satisfactory.

In the case when listening tests can't be performed or their setup is too expensive, the quality of transformed speech signals can be estimated using objective methods such as 3SQM or PESQ.

The 3SQM algorithm is applicable for speech quality predictions without a separate reference signal [3]. This method is recommended for non-intrusive speech quality assessment, live network monitoring and assessment by using unknown speech sources at the far-end side of a connection. The 3SQM approach is the first recommended method for single-ended non-intrusive measurement applications that takes into account the full range of distortions occurring in public switched telephone networks and that is able to predict the speech quality on a perception-based scale. The calculated score is then comparable to the quality perceived by a human listener, who is listening with a conventional shaped handset at this point.

PESQ (Perceptual Evaluation of Speech Quality) compares an original signal with a degraded signal that is the result of passing original signal through a communications system [1]. The output of PESQ is a prediction of the perceived quality that would be given to degraded signal by subjects in a subjective listening test. PESQ compares the original input signal with the aligned degraded output of the device under test using a perceptual model.

Performing subjective evaluations of digital codecs proceeds via a number of steps:

1. Preparation of source speech materials, including recording of talkers;
2. Selection of experimental parameters to exercise the features of the codec that are of interest;
3. Design of the experiment;
4. Selection of a test procedure and conduct of the experiment;
5. Analysis of results.

Speech material should consist of simple, short, meaningful sentences. These sentences should be chosen so as to be easy to understand (from the current non-technical literature or newspapers, for example). Further, the sentences should be made into sets of two or three in such a way that there is no obvious connection of meaning between the sentences in a set. Very short and very long sentences should be avoided, the aim being that each sentence when spoken should have a duration of 2–3 seconds. It is recommended that a minimum of two male and two female talkers should be used. However, if talker dependency is to be tested as a factor in its own right, it is recommended to use more talkers.

3. Characteristics of ITU-T Rec. P.50 Appendix I

This appendix to Rec. P.50 is a CD-ROM containing useful test signals. The signals on this CD-ROM include the signals described in Rec. P.50 as well as other signals that have been found useful by some administrations. Additionally, the full speech database that was used to develop Rec. P.50 is also on this CD-ROM.

Speech database consists of records of 20 languages and accents: English (American) and (British), Arabic, Chinese (Mandarin), Danish, Dutch, Finnish, French, German, Greek, Hindi, Hungarian, Italian, Japanese, Norwegian, Polish, Portuguese (Brazilian), Russian, Spanish (Castilian), Swedish. Each language (accent) is represented by 16 records (8 – male and 8 – female). Each record, of duration about 6–8 s, consists of 2–3 phrases. Records are written as 16 bits mono PCM files with 16 kHz sample rate. File names are in the form: `Lt_f1.wav`, where `Lt` – represents language (`Lt` – Lithuanian), `f` or `m` – female or male, correspondingly and `1` – record number.

4. Lithuanian Speech Records Database

The Lithuanian speech records database should not only satisfy requirements of ITU-T Rec. P.50 Appendix I, but also present the specifics of common Lithuanian language. At first, sentences should represent most frequently used language parts. Second, sentences must consist of most frequently used words.

According to research done in [15] and [16], the most frequent language parts in Lithuanian texts are nouns (∼45%), verbs (∼20%) and adjectives or pronouns (∼8%). The most frequently used nouns are the following: *galva, laikas, darbas, metas, vanduo, gyvenimas, motina, miestas, vakaras, širdis, pasaulis, saulė, dangus, vėjas*; verbs: *būti, galėti, nebūti, turėti, žinoti, eiti, sakyti, norėti, reikėti, matyti, žiūrėti, pasakyti, negalėti, imti, gyventi, atrodyti, ateiti, suprasti, stovėti, sėdėti, kalbėti*; and adjectives: *aukštas, valstybinis, administracinis, pagrindinis, politinis, bendras, naujas, didelis, tarptautinis, konstitucinis, ekonominis, atskiras, socialinis, visuomeninis*.

Basing on this research, four sentences consisting of 2–3 phrases were composed and recorded:

1. *Jonas buvo jaunas vyras, kuris moterims atrodė gražus, bet jo keistos kalbos jos negalėjo suprasti* (`Lt_f1.wav`, `Lt_f2.wav`, `Lt_m1.wav`, `Lt_m1.wav`).
2. *Taip galėjo sakyti tik žmogus, kurio širdis žinojo, kad šiame gyvenime galima pasitikėti tik savo tėvu* (`Lt_f3.wav`, `Lt_f4.wav`, `Lt_m3.wav`, `Lt_m4.wav`).
3. *Vakare aplink jį susirinkę stovėjo vyrai, tarytum atėję iš tuščio pasaulio, kuriame nebuvo saulės ir dangaus* (`Lt_f5.wav`, `Lt_f6.wav`, `Lt_m5.wav`, `Lt_m6.wav`).
4. *Po sunkaus darbo mieste motinai skaudėjo galvą, tačiau jos veidas visada buvo gražus, o balsas gyvas* (`Lt_f7.wav`, `Lt_f8.wav`, `Lt_m7.wav`, `Lt_m8.wav`).

In order to verify quality of recorded samples, we employed 3SQM quality estimation algorithm that is applicable for speech quality predictions without a separate reference signal [3]. We tested two Lithuanian database versions: the candidate (in this article denoted as LT1) recorded using high quality recording equipment and microphone, and the reference database (here denoted as LT2) recorded with built in laptop sound input device. LT2 has significantly lower signal to noise ratio then LT1. For comparison purposes, the same test was carried out on American English, British English and Arabic languages records taken from ITU-T Rec. P.50 Appendix I database. Obtained results (Fig. 1) show that individual and average quality estimates of records of all tested languages vary in a wide range. So 3SQM is not appropriate for absolute quality rating, however in the case of Lithuanian language we can state that LT1 version is a better candidate than LT2.

Further, to prove the applicability of our proposed Lithuanian speech records database for evaluating

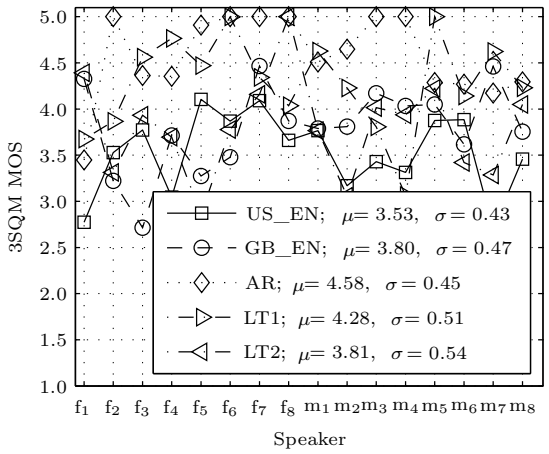

Figure 1. 3SQM scores for proposed Lithuanian (LT1 and LT2) records database as well as American (US), British (GB) English, Arabic (AR) speech records from ITU-T Rec. P.50 Appendix I database

quality of voice communication systems we experimentally tested the quality of two most popular voice codecs, AMR and Speex, which are widely used in telecommunications.

5. AMR and Speex Voice Codecs

AMR is an audio data compression scheme optimized for speech coding. AMR was adopted as the standard speech codec by 3GPP and is now widely used in GSM [13]. It uses link adaptation to select from one of eight different bit rates based on link conditions. The bit rates 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 and 4.75 kb/s are based on frames which contain 160 samples and are 20 ms long. AMR uses different techniques, such as: Algebraic Code Excited Linear Prediction (ACELP), Discontinuous Transmission (DTX), voice activity detection (VAD) and comfort noise generation (CNG).

Unlike many other speech codecs, Speex [14] is not targeted at cell phones but rather at voice over IP and file-based compression. The design goals have been to make a codec that would be optimized for high quality speech and low bit rate [14]. To achieve this, the codec uses multiple bit rates, and supports ultra wideband (32 kHz sampling rate), wideband (16 kHz sampling rate) and narrowband (telephone quality, 8 kHz sampling rate).

Speex is robust to lost packets, but not to corrupted ones since User Datagram Protocol ensures that packets either arrive unaltered or don't arrive. All this led to the choice of Code Excited Linear Prediction (CELP) as the encoding technique to use for Speex. One of the main reasons is that CELP has

long proved that it could do the job and scale well to both low bit rates and high bit rates.

6. Experimental Study

For the experiment we also employed both composed Lithuanian speech records databases (LT1 and LT2) and for comparison English (American and British) and Arabic speech records database from ITU-T Rec. P.50 Appendix I. During tests codecs were used in narrow band mode and set to operate at various bitrates. As bitrates of AMR codec are fixed, we also setup Speex codec to work at the same bitrates as AMR codec. The quality of encoded and decoded speech was estimated using PESQ algorithm. MOS scores for AMR and Speex codecs are depicted in Figs. 2 and 3, respectively.

From Fig. 2 it can be seen that the average quality estimates for speech records of all tested languages are close. Difference in the range of 0.1 MOS points is statistically insignificant. From Fig. 3 it is noticeable that Speex codec performed at lower quality than AMR on all tested languages, and speech with less signal to noise ratio (LT2) is more degraded than in AMR case. Also Speex codec shows less stability at lower bitrates. Difference in MOS points is greater than 0.5.

Furthermore, as it was already noted in [10], both codecs perform more stable on English language. In addition, in the case of Speex codec, mean MOS values for English language are about 0.2 points greater.

Also, it must be noted that, using Lithuanian records database LT1, both codecs produced speech of better quality (greater MOS values) than with LT2 database. So LT1 database should be considered as good match for ITU-T Rec. P.50 Appendix I database.

Lastly, to make sure that obtained results for Lithuanian database LT1 are correct, we performed subjective listening test by auditory from six persons whose native spoken language is Lithuanian. During this test, all listeners individually estimated quality (in MOS scale) of degraded speech record by comparing it with the original one. Quality estimates from all six listeners were averaged to produce MOS. Obtained results are shown in Fig. 4. This test also confirms that AMR codec performs better than Speex and quality estimates have very close tendency to the one obtained by using PESQ algorithm.

7. Conclusions

The results of the experiments show that:

- Created and extensively tested Lithuanian records database LT1 can be considered to be appropriate for inclusion into the ITU-T Rec. 50 Appendix I database.
- Language has influence on performance of tested voice codecs. Their stability of quality of reproduced speech significantly decreases at lower bitrates when coding speech signals of non-English languages. Furthermore, there also exists a noticeable bias to quality of reproduced speech of English language.
- In narrow band mode, AMR codec exhibits better and more stable performance (in the sense of quality of reproduced speech) than open source Speex codec.

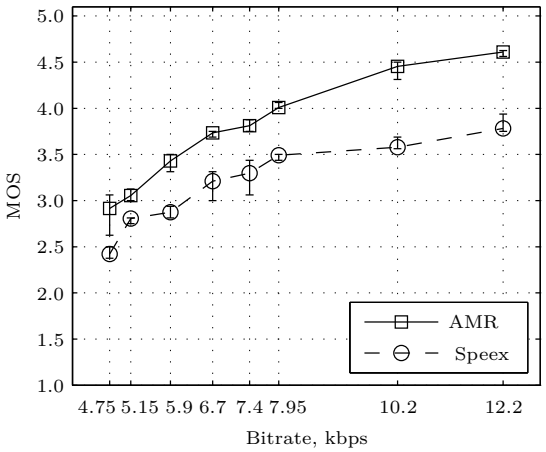

Figure 4. MOS scores for AMR and Speex codecs obtained during auditory listening test

Acknowledgements

Research in part was supported by Lithuanian State Science and Study Foundation project Reg. No. T-09022 and National Science Fund of Ministry of Education and Science of Bulgaria project Ref. No. DO-02-135/2008.

References

[1] Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862, 2001.

[2] Performance Characterization of the AMR Speech Codec (Release 5).3GPP TR 26.975 V5.0.0, 2002.

[3] Single ended method for objective speech quality assessment in narrowband telephony applications. ITU-T Recommendation P.563, 2004.

[4] Subjective performance assessment of telephone-band and wideband digital codecs ITU-T Recommendation P.830, 1996.

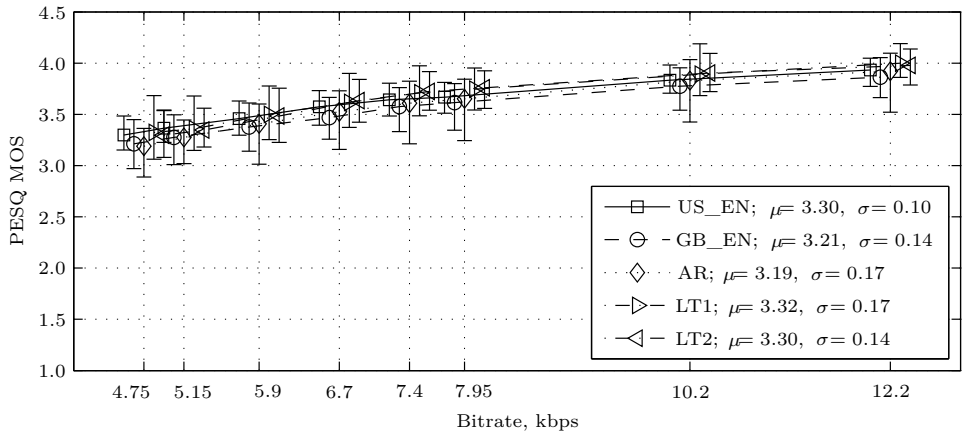

Figure 2. Mean and dispersion of PESQ MOS scores for AMR codec

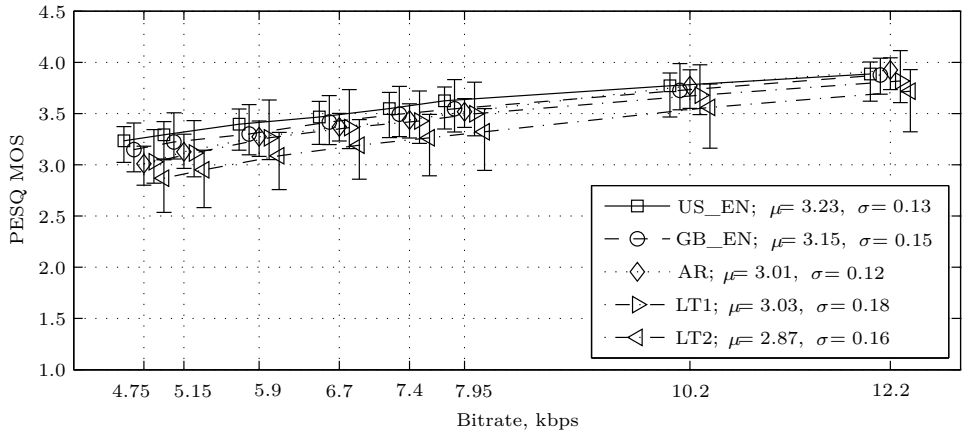

Figure 3. Mean and dispersion of PESQ MOS scores for Speex codec

[5] Telephone transmission quality, telephone installations, local line networks. Objective measuring apparatus: Test signals ITU-T Recommendation P.50 Appendix I, 1998.

[6] **I. S. Burnett and J. J. Parry.** On the effects of accent and language on low rate speech coders. In *Proceedings of Fourth International Conference on Spoken Language Processing, ICSLP 96*, 1996, vol. 1, pp. 291–294.

[7] **J. R. Deller, J. H. L. Hansen, and J. G. Proakis.** *Discrete-Time Processing of Speech Signals.* IEEE Press, 2000.

[8] **D. G. Jamieson, V. Parsa, M. C. Price, J. Till.** Interaction of Speech Coders and Atypical Speech I: Effects on Speech Intelligibility. *Journal of Speech, Language, and Hearing Research*, 2002, vol. 45, 482–493.

[9] **P. Ojala, H. Toukomaa, T. Moriya, and O. Kunz.** Report on the MPEG-4 Speech Codec Verification Tests. Technical report, MPEG Audio and Test subgroups, 1998.

[10] **M. Itani, S. Paulikas.** Influence of languages on CELP codecs performance. *Information Technology and Control*, 2008, vol. 37, no. 2, 141–144.

[11] **J. J. Parry, I. S. Burnett, and J. F. Chicharo.** Language-specific phonetic structure and the quantisation of the spectral envelope of speech. *Speech Communication*, 2000, vol. 32, no. 4, 229–250.

[12] **M. Schroeder and B. Atal.** Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'85*, 1985, vol. 10, pp. 937–472.

[13] **A. Uvliden, S. Bruhn, and R. Hagen.** Adaptive multi-rate: A speech codec adapted to cellular radio network quality. In *Proceedings of 32nd ASILOMAR Conference*,1998, vol. 1, pp. 343–347.

[14] **J. M. Valin.** *The Speex Codec Manual (version 1.0.4)*, 2004.

[15] **V. Žilinskienė.** Statistical analysis of the morphology of lithuanian administrative and publicistic styles. Lituanistica, 2002, vol. 49, no. 1, 106–116. (in Lithuanian)

[16] **V. Žilinskienė.** The use of grammatical forms in lithuanian works of fiction. Lituanistica, 2003, vol. 55, no. 3, 75–84. (in Lithuanian)

Received June 2009.

DOI: 10.5755/j01.itc.39.1.12092