2015

TCD-VoIP, a Research Database of Degraded Speech for Assessing Quality in VoIP Applications

Andrew Hines
Technological University Dublin, andrew.hines@tudublin.ie

Naomi Harte
University of Dublin, Trinity College

Eoin Gillen
University of Dublin, Trinity College

Recommended Citation

Hines, A., Harte, N. & Gillen, E. (2015). TCD-VoIP, a Research Database of Degraded Speech for Assessing Quality in VoIP Applications, *7th International Workshop on Quality of Multimedia Experience,* QoMEX, 2015. doi:10.1109/QoMEX.2015.7148100

TCD-VoIP, a Research Database of Degraded Speech for Assessing Quality in VoIP Applications

Naomi Harte, Eoin Gillen
Sigmedia Group,
School of Engineering,
Trinity College Dublin, Ireland

Andrew Hines
School of Computing,
Dublin Institute of Technology,
Ireland

***Abstract*—There are many types of degradation which can occur in Voice over IP calls. Degradations which occur independently of the codec, hardware, or network in use are the focus of this paper. The development of new quality metrics for modern communication systems depends heavily on the availability of suitable test and development data with subjective quality scores. A new dataset of VoIP degradations (TCD-VoIP) has been created and is presented in this paper. The dataset contains speech samples with a range of common VoIP degradations, and the corresponding set of subjective opinion scores from 24 listeners. The dataset is publicly available.**

I. Introduction

Voice-over-Internet Protocol (VoIP) refers to a group of technologies that allow users to communicate by voice over the internet. VoIP is expected to largely replace the legacy Public Switched Telephone Network (PSTN) by 2020 [1], [2]. Popular VoIP services include Microsoft Skype, Google Hangouts, and Apple FaceTime. VoIP offers many benefits over PSTN [3], one being that VoIP can employ more advanced algorithms and more bandwidth to deliver higher-quality speech. There are many issues which can affect speech quality as a whole, as discussed extensively by Möller et al. in [4]. For VoIP specifically, Karapantazis and Pavlidou [3] identified the three main parameters which affect the Quality of Service (QOS) as delay, jitter and packet loss rate.

Speech quality is a subjective concept. It is usually related to the intelligibility and "pleasantness" of the speech heard by a participant in the call [5]. Ideally, the quality of a speech sample is gauged by obtaining the subjective opinions of a statistically significant number of human listeners. For a listening-only test of transmitted speech with a wide range of impairments, the recommended test method is the Absolute Category Rating (ACR) test introduced in ITU Recommendation P.800 [6]. However, subjective listening tests require significant time and resources. It is impractical to run tests for every combination of codec, environment and impairment which occurs in VoIP. Thus, an automated, objective method of estimating subjective speech quality is desirable.

Over the last two decades, a range of algorithms have been proposed for this purpose [7]. The algorithms typically attempt to predict the MOS [6] of a subjective listener test on the data. Four such algorithms used for VoIP applications include: P.563 [8], PESQ [9], POLQA [10] and ViSQOL [11]. To develop or compare the performance of such algorithms, speech data with known MOS scores is needed. These speech samples should contain examples of quality issues which occur in VoIP independently of the hardware, network or codec in use. In effect, the degradations should be platform-independent.

Of the speech datasets which have been created for quality tests, very few include subjective MOS results. Of those that do, most have not been made available, and of the few that have, none contain the specific degradations salient in the VoIP scenario. Also, since VoIP is more commonly wideband [12], the samples should ideally represent the full bandwidth of audible frequencies (20-20000Hz). The NOIZEUS database [13] contains speech samples affected by noise along with subjective scores, but the speech has been downsampled to 8kHz. The ITU-T P Suppl. 23 database [14] contains speech affected by noise, packet loss and various codecs, as well as subjective scores. These speech samples have been downsampled to 16kHz. Other datasets exist (for example, the myriad of proprietary datasets used to test the POLQA objective metric [10]), but most are not publicly available.

Thus we were motivated to create the TCD-VoIP dataset. The categories mentioned above include platform dependent and independent degradations. For this database, five types of platform independent degradation were identified and chosen: background noise, intelligible competing speakers, echo effects, amplitude clipping, and choppy speech. Section II of the paper motivates the choice of the degradations used in TCD-VoIP. Section III explains how the parameters for each degradation were chosen. Section IV details the procedure used to run the subjective tests. Section V discusses the results, highlighting the useful attributes of this new dataset.

II. Degradation Choice

A. Background Noise

Noise has been a focus of several comparative studies. Falk and Chan [15] made use of car noise, Hoth noise and babble noise, while comparing PESQ to two other objective quality models. Kitawaki and Yamada [16] used subway, car, babble and exhibition noise while comparing PESQ with subjective results for noise-reduced speech. Hu and Loizou [13] used 8 types of noise in developing the NOIZEUS corpus of noisy speech data for evaluating speech enhancement algorithms. The types included "suburban train noise, multi-talker babble, car, exhibition hall, restaurant, street, airport and train-station noise". The noise was taken from the AURORA database, developed by Hirsch et al. [17]. We decided to use 4 types of commonly-occurring noise in TCD-VoIP: car, street, office and babble noise. The AURORA database has examples of these

978-1-4799-8958-4/15/$31.00 ©2015 IEEE

but the bandwidth is low. The European Telecommunications Standards Institute (ETSI) also created a database of noise samples as part of their study into speech quality in the presence of background noise [18]. They are available online and their bandwidth is 48kHz. Three noise samples from this dataset were used as the car, street and office noise samples. Krishnamurthy and Hansen [19] distinguished three categories of babble noise based on the number of speakers: competing speaker; babble; and large-crowd. Competing speaker was defined as having only two speakers. Babble had more speakers, but few enough that individuals and words were still occasionally identifiable. Large-crowd noise had enough speakers that no speaker or words could be identified. In tests, subjects rated all samples with more than 7 speakers as large-crowd noise. This is the desired type of noise for the babble samples in TCD-VoIP.

B. Competing Speakers

Since the speech in the competing speaker case is intelligible, we decided to treat it as a separate case to large-crowd babble. As explained by Krishnamurthy and Hansen [19], real background noise is usually made up of conversations. In a conversation, both speakers rarely speak at the same time. As a result, the competing speaker samples were all created using two speakers, speaking one after another. Since a two-person conversation may have two male speakers, two female speakers, or one male and one female, all three of these scenarios were represented.

C. Echo Effects

Echo effects in a voice call usually occur due to transmitted speech being picked up in the receiving unit's microphone, creating a feedback loop. However, echo effects can also be caused by other hardware issues. Strategies such as echo cancellation [20], where multiple delayed versions of the signal played at the receiving end are subtracted from the signal being returned, can be used to mitigate echo effects, but these are not completely effective. Mitigating functions also create their own effects on the signal. The most basic echo scenario was chosen for this database, where one or more copies of the transmitted signal are picked up by the receiving microphone and added to the returning signal. Since hardware or codec issues were not considered, the signal copies were simply attenuated and added, and not degraded in any other way. ITU-T Recommendation G.131 [21] offers guidance on how to mitigate talker echo in transmission systems. The recommendation contains a graph of echo delay against echo loudness relative to the original signal highlighting the "Acceptable" and "Limiting case" This guideline was used to inform choices of delay and loudness for the echo conditions in the dataset.

D. Clipping Effects

Clipping occurs when the amplitude of samples in a signal is set above the maximum permitted value. This causes the amplitude of those samples to be set at the maximum value (i.e. "clipped"), introducing distortions to the signal. Clipping can occur due to amplitude changes in hardware or software. It often occurs simply due to a speaker speaking too loudly into their microphone. The clipped speech samples used in this study are simply speech samples in which the amplitude has been raised by some constant, causing some proportion of the samples to be clipped.

E. Choppy Speech

Choppy speech in the context of VoIP refers to speech which is affected by missing samples. The most common cause of missing samples is packet loss in the VoIP network. Mitigation strategies such as Packet Loss Concealment (PLC) [22] can be used to smooth the effects of these missing samples. A number of studies have examined the effect of packet loss on speech quality [23], [24], [15], [25], [26]. However, since most are concerned with packet loss in the VoIP network, this is usually simulated by encoding the speech using a lossy VoIP codec and causing packet loss by running the stream through a virtual network. As a result, speech quality is affected by the packet loss and the codec. The focus of this study was on effects which occur independently of the codec or network. The effect of interest was loss of audio samples due to hardware overload. This can occur for example in a smartphone whose CPU is overloaded during a VoIP call, which can cause samples to be lost. These samples may be replaced by silence, or the previous samples repeated, or they may be skipped entirely. An example of this behaviour was encountered by Davies et al. [12] while testing the iSAC codec on an iPhone. They noticed "a rhythmic click (or chopping)", and attributed it to a lack of CPU capacity on the iPhone. This periodically choppy speech was the type chosen for this dataset.

III. Generation of Degraded Speech samples

A. Preparation of Source Material

The subjective test run in this study is the ACR test described in ITU-T Rec. P.800 [6]. Five ACR tests were run, one for each type of degradation. Rec. P.800 contains specific instructions for every aspect of the test. Instructions are given on suitable speech material, how this material should be recorded, how the talkers should speak, as well as other details. The TSP speech database from McGill University in Canada [27] was designed to offer researchers a common speech database to use for various experiments. Its speech material is consistent with the instructions in Rec. P.800. The dataset was recorded in an anechoic chamber and consists of 23 speakers reading sentences from the Harvard test sentence list. Each speaker recorded 60 sentences (though there are two sentences missing from speaker FE). The average length of each sentence is 2.4s.

In accordance with ITU-T Rec. P.800, the speech samples for the tests were created using sentences from four speakers (2 male, 2 female) from the TSP speech database. The ID codes of the speakers in question are FA, FG, MK and ML. Each speech sample consists of two sentences from a speaker, separated by a gap of two seconds, as well as a second of silence at the beginning and end of the sample. To avoid using the same sentences in the same order in each ACR test, each speaker's sentences were shuffled and new source speech samples were created for each test. The format of the resulting speech samples is shown in Table I.

TABLE I: Format of Speech samples

Silence (1s)	Sentence 1	Silence (2s)	Sentence 2	Silence (1s)

Once created, degradations were applied to the samples, after which their levels were normalized to -26dBov, as recommended by Rec. P.800 [6].

B. General Information for All Test Material

Each ACR test set comprised a number of conditions in which a varying amount of the relevant quality degradation was applied. The conditions are numbered (i.e. Condition 1, 2, 3 etc) with a higher number (usually) indicating a more severe degradation (except for the last 4 conditions - see Section III-C). In each test, Condition 1 is always a reference condition to which no degradation has been applied. Each condition is applied to a speech sample from all four speakers, meaning that if a test has 10 conditions, it has 40 degraded speech samples, one from each speaker for each condition. The intended function of the test material was for it to cover the full range of MOS values from 1-5. To achieve this, informal tests were done by the authors and a small group of volunteers. These tests guided the range of parameter choices for each condition.

In addition to the test samples, a set of practice samples containing one example of each condition was also created. The practice samples were created using sentences from the four test speakers, but the sentences are separate to those used in the test samples. Rec. P.800 [6] recommends that test participants be introduced to the range of degradations in the test before they begin rating test samples. The practice samples were played to participants before the test samples. Each participant heard the practice samples in the same randomized order. The speech samples in the TSP speech database are 16-bit WAV files sampled at 48kHz. Since the purpose of these tests was to examine degradations independently of codec or channel effects, this is also the final format of the output speech samples. A summary of the degradations and parameters used is given in Table II. Complete details on all conditions are detailed in the database documentation, downloadable from www.mee.tcd.ie/˜sigmedia/Resources along with all the complete dataset.

C. Modulated Noise Reference Units (MNRUs)

A Modulated Noise Reference Unit (MNRU) is defined in ITU-T Rec. P.810 [28] as a standalone unit that is intended to introduce controlled degradations to speech signals. ITU-T Rec. P.800 [6] suggests that it may be appropriate to include MNRUs in ACR tests as reference conditions. Their purpose is to enable scores in different tests to be compared. Since there were five subjective ACR tests to be conducted in this work, involving different types of degradation, MNRUs were used. This allowed results from the tests to be compared. All tests bar the ”Speech with Echo Effects” test (see Section III-F) include MNRUs.

The process of creating an MNRU version of a speech sample is detailed in ITU-T Rec. P.810 [28]. Essentially, the process involves adding speech-shaped noise to the signal at a desired signal-to-noise ratio, called the Q factor. The noisy output is low-pass filtered at 7kHz as a postprocessing stage. The ITU provides software on their website [29] which can be compiled and used to create MNRUs but it is designed to work with input audio sampled at 16kHz. The sampling rate used for all speech samples in the tests is 48kHz. Hence, the workaround used was to downsample the speech samples for which MNRUs were to be made to 16kHz. The downsampled versions were passed into the ITU's MNRU program. The output MNRUs were then upsampled back to 48kHz.

MNRUs at four Q levels were added to the Background Noise, Competing Speaker, Choppy Speech and Clipped Speech ACR tests. The Q levels used were 48, 36, 24 and 12. These levels were previously used by GIPS and Cisco in their tests of the iSAC VoIP codec. Since there were 4 speakers, an MNRU at each Q level was made for each speaker, leading to a total of 16 MNRUs to include in each test. The MNRUs are treated as extra conditions for each degradation type, and always have the four highest condition numbers (e.g. in the choppy speech test, the MNRUs have the condition numbers 21, 22, 23 and 24).

D. Speech with Background Noise

Four types of noise (see Section II-A) were tested: speech babble noise, car noise, road noise and office noise. The car, road and office noise samples were sourced from the ETSI noise database [18]. The speech babble sample was created specifically for this test by combining random sentences from 10 random speakers from the TSP speech database. All of the speakers and sentences in the babble noise were separate to those used to create speech samples. There were two varying parameters in this test: the type of noise (babble, car, road or office) and the SNR of the noise. 20 combinations of these two parameters were used in the test.

E. Speech with Competing Speakers

The scenario in this test is that the main speaker has to compete with a conversation which is ongoing at some distance from them (see Section II-B). Since the scenario envisions a conversation in the background, we decided to mimic this by playing a random sentence from a speaker in the TSP Speech database, then playing a random sentence from a second speaker, then another random sentence from the first speaker, etc. There are limits on which speakers and sentences can be used for this purpose. The four speakers (FA, FG, MK, ML) being used as the source material cannot be background speakers, and the same sentence cannot be used twice in one sample.

Using pairs of background speakers to generate background “conversations” also introduces scenarios where both background speakers are female, both are male and where one is male and one female. As a result there were two parameters varied in this test: the Signal-to-Noise Ratio (SNR) of the competing conversation, and the genders of the participants (coded as ff=both participants are female, mm=both participants are male, and mf/fm=one male and one female participant). 10 combinations of these parameters were used in the test. To differentiate the intended speaker from the competing speaker, the intended speaker begins speaking 500ms before the competing speaker in all samples. This is explained to listeners before the test.

F. Speech with Echo Effects

The echo effect tested here (see Section II-C) was produced in the speech samples by adding one or more delayed versions of the signal to the original signal at specified SNRs. Three parameters were varied in this test:

1) Echo Alpha: Amplitude (%) of the first delayed version of the signal relative to the original.
2) Echo Delay: Delay (in ms) of first delayed version of the signal relative to the original.
3) Feedback Factor: Amplitude (%) reduction to apply to each subsequent echo to emerge from the echo feedback loop.

20 combinations of the three parameters above were used in this test.

G. Clipped Speech

There was only one varying parameter in this test: the amplitude multiplier applied to the samples (see Section II-D). The amplitude of the speech is increased by multiplying each sample by the multiplier. The maximum and minimum permitted values in a WAV file are plus and minus 1 respectively. Any samples which exceed these values after increasing their amplitude are set (clipped) to plus or minus 1. 10 values of the multiplier parameter were used in this test.

H. Choppy Speech

Three parameters were varied to create the choppy speech samples:

1) Chop Mode: Determines whether samples will be replaced with zeroes (mode 1), deleted entirely (mode 2) or overwritten with the previous portion of samples (mode 3)
2) Chop Period: Determines the length (ms) of each portion of samples to be chopped.
3) Chop Rate: Determines how often (ms) a portion of samples should be chopped.

This modelled the three scenarios of Section II-E. 20 combinations of the three parameters were used.

TABLE II: Degradations and Parameters used in TCD-VoIP

Degradation	Conditions	Parameters	Range
Chop	20	Rate Period Mode	0-6 chops/s 0.02-0.04 s Insert, Delete, Overwrite
Clip	10	Multiplier	1-55
Competing Speaker	10	Gender code SNR	1-5 10-50 dB
Echo	20	Alpha Delay	0-0.5 0-220 ms
Noise	20	Noise Type SNR	Car, Street, Office, Babble 5-55 dB
MNRUs	4	SNR (Q)	48, 36, 24, 12

IV. ACR Test Procedure

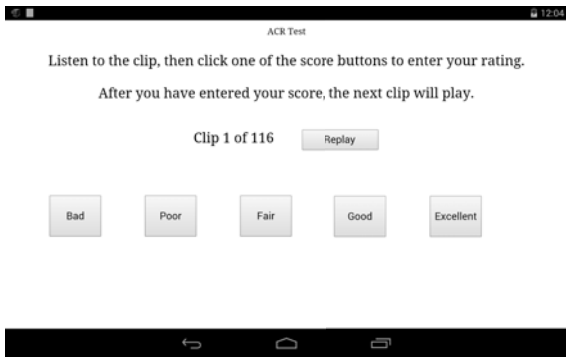
A. Listening Environment

The listening room used for the ACR tests is a soundproof recording studio. The background noise in the recording studio was measured at 30 dBA and reverberation time was 0.18s at 1kHz, spatially averaged. This satisfies the requirements of ITU-T Rec. P.800 [6] for the listening environment.

B. Listening System

The listening system used in the tests consisted of a Nexus 7 tablet computer (2013 edition) and a pair of Sennheiser HD558 headphones. The HD558s are high-quality, semi-open-backed headphones. The test software was delivered on the Nexus 7 using a web UI written in HTML5 and Javascript. To facilitate the loading of samples and storing of scores,

Fig. 1: ACR Test Software sample Scoring Screen



the webpage was hosted locally on the Nexus 7 itself, and accessed by way of a local server which was running on the Nexus 7. The local server app is called "BitWebServer". The test software's user interface is shown in Figure 1. The listener begins the test by pressing the "Start Test" button, after which the first pair of sentences plays. When the first sample is finished playing, the score buttons are revealed. At this point, the listener can either replay the sample, or select a quality score, at which point the score buttons are hidden while the second pair of sentences plays. The test proceeds in this manner until the end, where a "Test Finished!" screen is shown. The process of deactivating the score buttons until the current sample has finished playing prevented listeners rating a sample before they have heard it in its entirety.

C. Listening Level

In accordance with Rec. P.800 [6], a calibration tone was created. The calibration tone was generated in Matlab, and consisted of a 1kHz sine wave at a level of -26dBov for 30 seconds. It was used to set the volume of the playback system to 79dB SPL at the ear reference point. This level was measured with a sound level meter.

D. Listeners

24 listeners were used in all experiments. The 24 listeners used for the chop, clip, noise and background speaker tests consisted of 13 males and 11 females, while the 24 listeners used for the echo test consisted of 17 males and 7 females. The eligibility criteria for listeners were based on the criteria used by ITU-T Study Group 12 while testing the G.729 codec [14]: they had not previously been directly involved in work on speech quality assessment, they had no known hearing defects, they were native speakers of the language in which the tests were being conducted and they were between the ages of 18 and 65.

E. Procedure

Participants were seated in the studio and introduced to the testing hardware. The test itself was explained as follows:

"You are about to hear a number of pairs of sentences. After each pair of sentences, you will be asked to rate your opinion of their quality on a scale from Bad to Excellent. The meaning of 'quality' is left up to you, but typically it includes factors such as intelligibility and pleasantness".

For the "Speech with Competing Speaker" test (Section III-E), an additional instruction was given to explain which

speaker is the intended speaker and which are the competing speakers. The participant then began the test. After listening to some of the practice samples, the test assistant verified with the participant that they understood the test procedure. If so, the test assistant then left the studio so the participant could complete the test in silence.

F. Randomization of Test samples

ITU-T Rec. P.800 [2] recommends a number of methods which are suitable for randomizing the samples in an ACR test. The choice of method is left to the experimenter. The randomized block design used by ITU-T Study Group 12 [14] during their tests was followed for these tests. Under this design, samples are split into the same number of blocks as there are speakers (i.e. 4 blocks in this case). One speech sample for each condition must occur in each block. Each speaker must contribute the same number of samples to each block. Two samples from the same speaker may not occur in direct succession. Taking these constraints into account, the order of the samples in each block is randomized. For their tests, ITU-T Study Group 12 created four randomized orders of their samples using this design, and used each order on 6 listeners. The initialization routine of the test software allows a new randomized order for each listener. This accommodates a variable number of listeners. This approach was used on the samples for all tests.

G. Duration of Tests

Listeners completed the Background Noise, Competing Speaker, Choppy Speech and Clipped Speech tests in one visit. There were a combined total of 364 samples in these four tests. To rate all 364 samples in one test session would take approximately 75 minutes. ITU-T Rec. P.800 [6] recommends that ideally a test session should last no more than 20 minutes, and that no listener should rate samples for more than 45 minutes without a break. To comply with this directive, the tests were split into two sessions of approximately 37 minutes each. Listeners were given a 10-minute break between the two sessions for refreshments. The "Speech with Echo Effects" test was completed separately to the other four tests. It contained 100 samples and took each listener 20 minutes on average to complete.

V. Results

Figure 2 shows the results of the five subjective tests. It can be seen that the conditions in each degradation test represent a wide range of MOS values from 1–5, which was the desired outcome. The reference condition in each test has obtained a score of roughly 4.5, while the most severe conditions have obtained scores of 1.5 or slightly higher. The reference MNRU conditions (visible as the last 4 conditions in white in Figures 2a, 2b, 2d and 2e) have also obtained consistent scores across the four tests. This shows that the listeners were rating in a consistent way, and allows the other scores from the tests to be compared.

It is informative to note how severe degradations were before subjects gave a MOS score of 3 or lower. For background noise, only conditions where the noise SNR was below 20dB yielded MOS less that 3. Babble noise was consistently more annoying. In the competing speaker scenario, the gender of the competing speaker was found to make no significant difference. The dominant factor was the SNR of the competing speaker which needed to be below 20dB to give a MOS lower than 3. For echo, conditions 15,16, 18 and 20 which included feedback, yielded the lowest MOS. This is a commonly encountered issue in VoIP and this dataset highlights the impact feedback has on quality. In the absence of feedback, even with a long delay at 200ms in condition 7, because the alpha value was only 3%, the MOS remained high. In contrast, condition 17 has a delay of 180ms but an alpha value of 30%, yielding a MOS of below 2. It is worth noting that the standard parametric echo estimation from the E-model will not consider feedback factor.

Clip gain had to exceed a multiplier of 18 before MOS dropped below 3 for the clip conditions. The multipliers in condition 8, 9 and 10 are 18, 25 and 55 respectively. Thus condition 10 represents a situation where 80% of all samples were clipped which would not typically occur in VoIP. For chop, with fixed period and rate, the mode of chop significantly influences the score. Condition 9, 13 and 17 all have a chop period of 40ms and a chop rate of 2 sec. Repeating, rather than deleting or inserting silence gave the lowest MOS score. Chop rates had to exceed 3s before MOS scores were consistently below 3. Beyond this chop rate, deletion yielded the best MOS scores.

One limitation of this data is that degradations have been treated in isolation. In a live VoIP call, a users quality of experience may typically be impacted by all or some of these effects. It is still useful for developers to have an insight into how annoying individual degradations can be, as these MOS scores reveal that users are actually very tolerant of these degradations when their level is low.

VI. Conclusion

TCD-VoIP is a freely available dataset of degraded speech samples with corresponding subjective opinion scores. The range and level of degradations makes this database a useful resource for the development and testing of speech quality metrics in VoIP scenarios. We hope the data will allow direct comparison of how different quality metrics perform for VoIP. The database and full supporting documentation are available at www.mee.tcd.ie/˜sigmedia/Resources.

Acknowledgment

The authors would like to thank Google Inc. for sponsorship of this research.

References

[1] Alcatel-Lucent, "PSTN industry analysis and service provider strategies: Synopsis," http://goo.gl/tTPFes, Alcatel-Lucent, Paris, France, Tech. Rep. Bell Labs Analysis for BT, 2013.

[2] L. K. VANSTON and R. L. HODGES, "Forecasts for the us telecommunications network," *Telektronnik*, vol. 104, no. 3/4, pp. 18–28, 2008.

[3] S. Karapantazis and F.-N. Pavlidou, "Voip: A comprehensive survey on a promising technology," *Comput. Netw.*, vol. 53, no. 12, pp. 2050–2090, aug 2009. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2009.03.010

[4] S. Möller, F. Köster, F. Schiffner, and J. Skowronek, "Analyzing technical causes and perceptual dimensions for diagnosing the quality of transmitted speech," *4th International Workshop on Perceptual Quality of Systems, Vienna*, pp. 30–35, 2013.

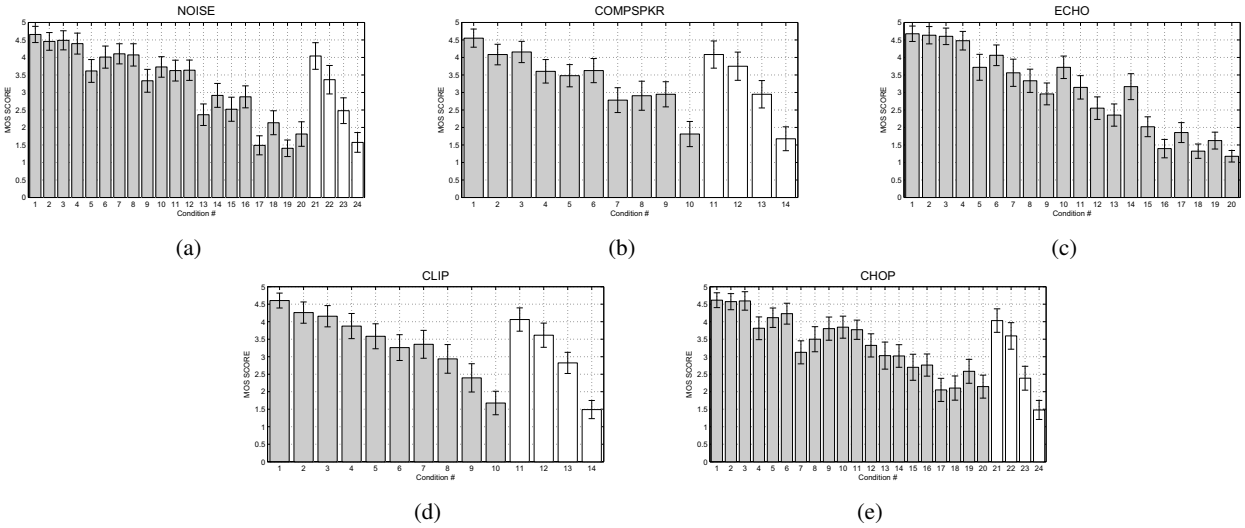


Fig. 2: The subjective MOS results from the five degradation types. The MOS value for a condition is the average score given by all listeners to the four speech samples affected by that condition. The error bars are 95% confidence intervals obtained using the method in ITU-T Rec. P.1401 [30]. As mentioned in Section III-B, the 1st condition in each figure represents an undegraded reference condition. The MNRUs are highlighted in white. (see Section III-C).

[5] K. Kondo, *Subjective Quality Measurement of Speech*. Springer, 2012.

[6] Int. Telecomm. Union, "Methods for subjective determination of transmission quality," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.800, 1996.

[7] S. Moller, W.-Y. Chan, N. Cote, T. Falk, A. Raake, and M. Waltermann, "Speech quality estimation: Models and trends," *Signal Processing Magazine, IEEE*, vol. 28, no. 6, pp. 18–28, Nov 2011.

[8] Int. Telecomm. Union, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.563, 2004.

[9] ——, "Perceptual evaluation of speech quality (pesq): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.862, 2001.

[10] ——, "Perceptual objective listening quality assessment," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.863, 2011.

[11] A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The Virtual Speech Quality Objective Listener," in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, Sept 2012, pp. 1–4.

[12] M. Davies, D. Elwood, and C. Politis, "Implementing a superwideband codec for smartphone voip services," in *Wireless Conference (EW), Proceedings of the 2013 19th European*, April 2013, pp. 1–6.

[13] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, pp. I–I.

[14] Int. Telecomm. Union, "ITU-T coded-speech database," http://handle.itu.int/11.1002/1000/4415, Feb 1998.

[15] T. H. Falk and W.-Y. Chan, "Performance study of objective speech quality measurement for modern wireless-voip communications," *EURASIP J. Audio Speech Music Process.*, vol. 2009, pp. 12:1–12:11, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/104382

[16] T. Yamada, M. Kumakura, and N. Kitawaki, "Subjective and objective quality assessment of noise reduced speech signals," in *Nonlinear Signal and Image Processing, 2005. NSIP 2005. Abstracts. IEEE-Eurasip*, May 2005, pp. 28–.

[17] D. Pearce, H.-G. Hirsch, and E. E. D. Gmbh, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *in ISCA ITRW ASR2000*, 2000, pp. 29–32.

[18] European Telecomm. Standards Union, "Speech quality performance in the presence of background noise - part 1: Background noise simulation technique and background noise database," ETSI, Sophia-Antipolis Cedex, France, Tech. Rep. ETSI EG 202 396-1, 2008.

[19] N. Krishnamurthy and J. Hansen, "Babble noise: Modeling, analysis, and applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1394–1407, Sept 2009.

[20] M. Sondhi, "An adaptive echo canceller," *Bell System Technical Journal*, vol. 46, no. 3, pp. 497–511, 1967.

[21] Int. Telecomm. Union, "Talker echo and its control," http://www.itu.int/rec/T-REC-G.131/en, Nov 2003.

[22] H. Sanneck, A. Stenger, K. Ben Younes, and B. Girod, "A new technique for audio packet loss concealment," in *Global Telecommunications Conference, 1996. GLOBECOM'96.'Communications: The Key to Global Prosperity*. IEEE, 1996, pp. 48–52.

[23] S. Jelassi and G. Rubino, "A study of artificial speech quality assessors of voip calls subject to limited bursty packet losses," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 1, p. 9, 2011. [Online]. Available: http://jivp.eurasipjournals.com/content/2011/1/9

[24] V. A. Reguera, F. F. Á. Paliza, W. Godoy, and E. M. G. Fernández, "On the impact of active queue management on voip quality of service," *Computer Communications*, vol. 31, no. 1, pp. 73–87, 2008.

[25] J. Holub and O. Slavata, "Impact of IP channel parameters on the final quality of the transferred voice," in *Wireless Telecommunications Symposium (WTS), 2012*, April 2012, pp. 1–5.

[26] J. Turunen, P. Loula, and T. Lipping, "Assessment of objective voice quality over best-effort networks," *Comput. Commun.*, vol. 28, no. 5, pp. 582–588, mar 2005.

[27] P. Kabal, "TSP speech database," McGill University, Quebec, Canada, Tech. Rep. Database Version 1.0, 2002.

[28] Int. Telecomm. Union, "Modulated noise reference unit (MNRU)," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.810, 1996.

[29] ——, "Software tools for speech and audio coding standardization," http://www.itu.int/rec/T-REC-G.191/en, Aug 2014.

[30] ——, "Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," ITU, Geneva, Switzerland, Tech. Rep. ITU-T Rec. P.1401, 2012.