Lubica Blaskova – Jan Holub *

HOW DO NON-NATIVE LISTENERS PERCEIVE QUALITY OF TRANSMITTED VOICE?

This paper describes the test methodology and results of speech transmission quality testing in the environment of non-native listeners; it means the communication language differs from the mother tongue of the test subjects. The tests were carried out based on ITU-T P.800 on a database of English speech samples affected by various coding distortions and background noise conditions. The subjects were pre-tested on their English proficiency. The subjective test results confirm a systematic and repeatable shift in subjective quality assessment performed by non-native listeners.

1. Introduction

In many practical cases, the communication in the telecommunication network is carrying a non-native language for one or more conversation participants [1]. There are procedures to automatically estimate perceived quality of the transmitted speech [3]-[5] and their results correlate well with subjective experiments carried on native speakers and native listeners. However, it is not clear if the effect of listener non-nativity can affect the quality perception. This paper examines methods to quantify such effects by presenting listening test results performed on non-native listeners, pre-sorted according their English proficiency.

1.1. Speech Transmission Quality Measurement

Speech transmission during any call in the telecommunication network is affected by many impairments; including delay, echo, various kinds of noise, speech (de)coding distortions and artefacts, temporal and amplitude clipping etc. Each transmission impairment has a certain perceptual impact on the speech transmission quality. The overall quality can be evaluated and expressed in terms of a Mean Opinion Score (MOS) covering the range from 1 (bad) to 5 (excellent). Speech transmission quality measurements are widely used to compare different coding and transmission technologies, or to monitor the network performance. The traditionally proven but expensive subjective methods [2], involving human listeners assessing many speech samples, have been partially replaced by objective digital signal processing algorithm based measurements that either compare the original undistorted signal to the received one [3] (so called intrusive or double-sided algorithms) or process only the received version [5]. All these methods have been designed and tested on past and contemporary telecommunication transmission standards that are widely used in common mobile and fixed telecommunication networks, e.g. those using 'toll quality' voice encoding.

1.2. Non-nativity as a Quality Perception Factor?

In many important practical cases, the communication in the telecommunication network is carrying a non-native language for one or more conversation participants. Typical examples are e.g. international and/or roaming calls in today's public fixed and mobile telecommunication networks or communications in military radio telecommunication networks during multi-national tactical operations [6], [7], international governmental organisations or multi-national companies. As the objective methods should be as accurate as possible replacements of subjective methods, the question of the influence of non-nativity to the final quality perception arises. Unfortunately, there are contradictory hypotheses about such an influence:

- Non-native listeners have higher difficulties to understand the contents even for less distorted samples than native listeners, thus they should assess quality worse (=giving generally lower scores) than native listeners.
- Non-native listener's brain is more occupied by message content decoding than in case of native listener, thus the quality assessment should not be so detailed, so some impairments can be missed, thus the final scores should be higher than for native listeners.

2. Work performed

2.1 Selection of Coders and Database Recording

A speech database fulfilling P.800 requirements and containing two background noise conditions (no noise/Hoth noise +10dB SNR) were recorded on selected coders (PCM 8 bit, GSM 06.10, MELPe 2.4 kbit/s). The final database contained 120 different sentences spoken by native English speakers. More than 2 female and 2 male speakers recorded in studio environment were used. In each case, 15 sentences per condition (noise+coder, see Table 1)

* **Lubica Blaskova, Jan Holub**
Department of Measurement, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic,
E-mail: l.blaskova@mesaqin.com, holubjan@fel.cvut.cz

were prepared. The active speech level as per ITU-T P.56 was equalized to -26 dBoV that corresponded then to 79 dB SPL (A) during the listening tests.

2.2. Selection of Listeners and their Language Proficiency Testing

Subjective tests were carried out on naive subjects as required by P.800. Their age was in the range between 20 and 30. None of them was a native English speaker, the nationalities represented in the group were: Czech, Slovak, Italian. The English proficiency of each subject was verified by a short quiz consisting of played-out English sentences/articles and followed by a set of questions to be answered on a written/multiple-choice principle. The language test lasted 7 minutes and was always performed right before the quality testing. The maximum achievable number of points in the language test was 21. Based on the language test results, the subjects were assigned to one of 3 categories:

- Beginners (0-3 points)
- Intermediate (4-10 points)
- Advanced (11-21 points)

The subjects were not informed about their results after the language tests.

2.3. Subjective Testing

Subjective tests as per ITU-T P.800 [2] were performed on the 120 sample database as described in 2.1. The subjective listening-only tests were performed in a critical listening room where up to 8 listeners could be seated. The reverberation time of the room is 185 ms and natural background noise less than 10dB SPL (A). The samples were played-back in random order by means of a digital playback system with SNR higher than 105 dB. The loudspeakers were actively compensated to achieve transition ripple less than 0.8 dB in an audible frequency range.

Multiple sessions were run always with different listeners. In total, 36 votes per sample were obtained, 13 per Beginners, 11 per Intermediate and 12 per Advanced groups.

2.4. Subjective Testing on Native Listeners

For verification purposes, similar subjective tests as described in 2.3 were performed using the same database but on native listeners. This experiment was carried out in Los Angeles, California, in April 2008. The test subjects were students of California State Polytechnic University, Pomona. The purpose of the test was to compare influence of (non-) nativity and different expectations of both groups of subjects, coming at the same time from different continents. Test subjects were seated in standard class room with only basic anechoic measures (plasterboard lining). The play-out system used non-compensated loudspeakers with transition ripple up to 9 dB in an audible frequency range.

3. Results

Test results are given in the following tables and figures. A special attention was paid to differences in quality perception between the Advanced group and the remaining two non-native (Intermediate and Beginners) groups. The per-condition results are listed in Table 1 and shown in Figure 1. Figure 2 shows results per sample. Both per-condition and per-sample results showed clear shift in subjective scoring of non-native listeners and the difference between Advanced and both other groups (Beginners and Intermediate) is about 0.5 MOS for the entire MOS scale. The difference between the Intermediate and Beginners is not so evident (not shown in the figures) and fits within confidence intervals of subjective experiments. The results of native listeners testing are reported in Fig. 3.

Due to different expectations driven by different communication technologies used in different countries and also due to different environmental conditions (room and equipment) the experimental results achieved on native and non-native listeners can not be directly compared. This is also well noticeable from Table 2 where Pearson correlation coefficients are reported. The results coming from native listeners provide significantly lower correlations with all other listener groups than in the remaining rows (where the results between two non-native listener groups are reported). Note that the correlation calculation is invariant to offset and gain changes so the systematic offset identified between Advanced and other non-native groups does not influence the results in Table 2.

Subjective test results (per condition) Table 1

Condition	Noise type	Coder	MOS-LQSn Advanced	MOS-LQSn Intermed.	MOS-LQSn Beginners
1	no noise	clean	4.38	3.78	3.88
2	no noise	PCM 8bit lin.	3.32	2.87	3.09
3	no noise	GSM 06.10	2.51	1.81	1.75
4	no noise	MELPe 2.4	2.87	2.58	2.44
5	10 dB Hoth	clean	3.55	2.74	2.61
6	10 dB Hoth	PCM 8bit lin.	2.75	2.35	2.36
7	10 dB Hoth	GSM 06.10	1.94	1.48	1.33
8	10 dB Hoth	MELPe 2.4	2.16	1.91	1.85

Pearson correlation coefficients between different listener groups ("per condition" results) Table 2

	Native	Advanced	Intermediate	Beginners
Native	1.000	0.670	0.789	0.758
Advanced		1.000	0.972	0.957
Intermediate			1.000	0.991
Beginners				1.000

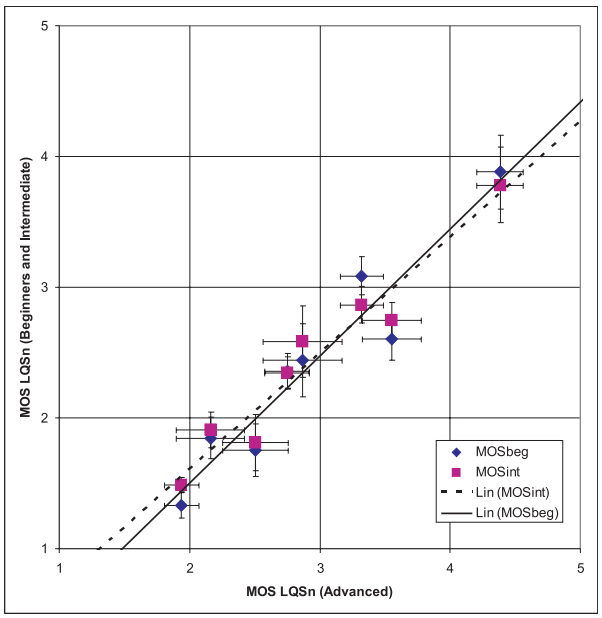

Fig. 1 Subjective test results per condition, comparison between Advanced and both other (Intermediate and Beginners) groups. 95% confidence intervals (CI95) are reported.

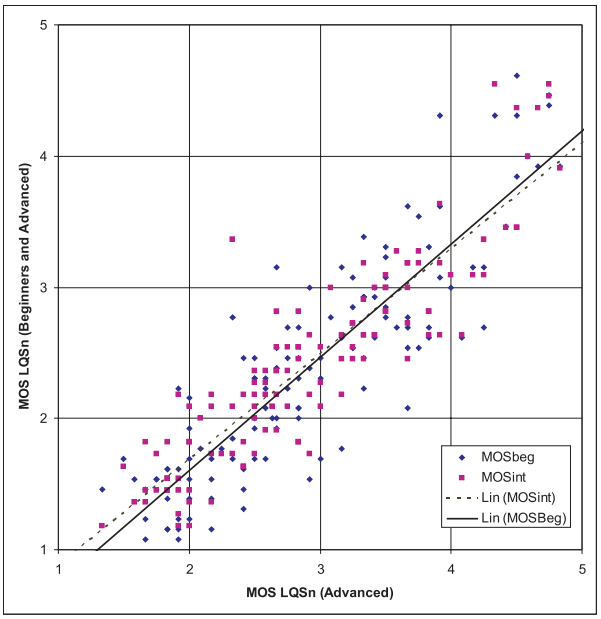

Fig. 2 Subjective test results per sample, comparison between Advanced and both other (Intermediate and Beginners) groups

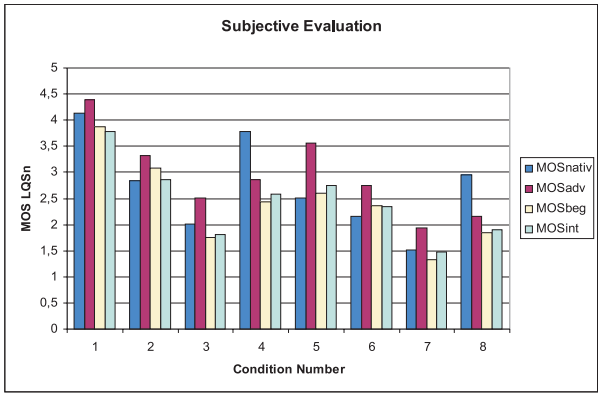

Fig. 3 Comparison between non-native and native listeners with different expectation factors

4. Conclusions

It is evident from the results that both non-advanced groups of non-native listeners (means Beginners and Intermediate) scored the samples systematically lower than Advanced listeners. It means that the first hypothesis listed in Chapter 1.2 was confirmed. The offset is approximately 0.5 MOS along the entire MOS scale.

This systematic offset can be conveniently used to re-map PESQ or other objective algorithm output to bring the algorithm result closer to "conventionally correct" (meaning subjective) results in case the communication in the telecommunication network is carrying a non-native language for one or more conversation participants which occur e.g. during international and/or roaming calls in today's public telecommunication networks or communications in military telecommunication networks during multi-national tactical operations. Such correction can impact significantly e.g. threshold-based decisions on link quality acceptability in automatic measurements performed by network monitoring systems or drive-test systems.

Acknowledgment

This work has been supported by the Czech ministry of Education: MSM 6840770014 "Research in the Area of the Prospective Information and Navigation Technologies". The authors would like to thank Dr. Koichiro R. Isshiki from Cal Poly Pomona and Dr. Michael Street from NATO C3A for their valuable help and assistance.

References

[1] STREET, M. D.: *NGN in the Military Domain: NATO Network Enabled Capability (NNEC) and Networked Information Infrastructure (NII)*, RTO LS-070, paper 10, 2007.

[2] ITU-T Rec. P. 800 *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva, 1996.
[3] ITU-T Rec. P. 862 *Perceptual Evaluation of Speech Quality*, International Telecommunication Union, Geneva, 2001.
[4] PENNOCK, S.: *Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) Algorithm*, MESAQIN 2002, Praha, CTU.
[5] ITU-T Rec. P. 563: *Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications,* International Telecommunication Union, Geneva, 2004.
[6] STREET, M, COLLURA, J.: *Interoperable Voice Communications: Test and Selection of STANAG 4591*, RTO-IST conf. on 'Military communications', Warsaw, Poland, 2001.
[7] HOLUB, J., STREET, M., SMID, R.: *Intrusive Speech Transmission Quality Measurements for Low Bit Rate Coded Audio Signals,* AES115 Convention, New York, October 2003.