How do Non-native Listeners Perceive Quality of Transmitted Voice?

Ľubica Blašková, Jan Holub

Dept. of Measurement K13138, FEE CTU Prague, Technicka 2, CZ 166 27 Prague 6, Czech Republic
tel. +420 2 2435 2131, fax: +420 2 2435 2199, holubjan@fel.cvut.cz

ABSTRACT

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 mother tongue of the test subjects. The tests have been carried out based on ITU-T P.800 on the database of English speech samples affected by various coding distortions and background noise conditions. The subjects have been pre-tested on their English proficiency. The subjective tests results confirm systematic and repeatable shift in subjective quality assessment performed by non-native listeners.

1.0 INTRODUCTION

In many practical cases, the communication in the telecommunication network is carried in non-native language for one or more conversation participants. 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 tests 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 [4]. 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 carried in 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]. As the objective methods should be as accurate as possible replacements of subjective methods, the question about influence of non-nativity to 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 as 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, means some impairments can be subjectively missed, thus the final scores should be higher than for native listeners.

2.0 WORK PERFORMED AND RESULTS

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) has been 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, have been used. Always 15 sentences per condition (noise+coder) has been prepared. The active speech level as per ITU-T P.56 has been 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 have been carried on naive subjects as required by P.800. Their age was in the range between 20 and 30. None of them was native English speaker, the nationalities represented in the group were: Czech, Slovak, Italy. The English proficiency of each subject has been verified by short quiz, composed by played-out English sentences/articles and followed by set of questions to be answered in written in 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] have been performed on the 120 sample database as described in 2.1. The subjective listening-only tests have been performed in a critical listening room where up to 8 listeners can be seated. The reverberation time of the room is 185 ms and natural background noise less than 10dB SPL (A). Multiple sessions have been run always with different listeners. In total, 36 votes per sample have been obtained, 13 per Beginners, 11 per Intermediate and 12 per Advanced groups.

3.0 RESULTS

Test results are given in the tables and figures. Special attention has been paid to differences between Advanced and remaining two groups quality perceptions. Per-condition results are listed in Table 1 and shown in Figure 1. Figure 2 shows results per sample.

4.0 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. This 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” (means subjective) results in case the communication in the telecommunication network is carried in non-native language for one or more conversation participants. Such correction can impact significantly e.g. threshold-based decisions on link quality acceptability in automatic measurement 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”.

REFERENCES

[1] Technical Note 1246, ‘Wireless communication architecture (land tactical):

scenarios, requirements and operational view', NC3A, The Hague, Netherlands, 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 and 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., Šmíd, R. : Intrusive Speech Transmission Quality Measurements for Low Bit Rate Coded Audio Signals, AES115 Convention, New York, October 2003

Table 1. Subjective test results (per condition)

Noise type	Coder	MOS-LQSn Advanced	MOS-LQSn Intermediate	MOS-LQSn Beginners
no noise	clean	4,38	3,78	3,88
no noise	PCM 8bit	3,32	2,87	3,09
no noise	GSM 06.10	2,51	1,81	1,75
no noise	MELPe 2.4	2,87	2,58	2,44
10 dB Hoth	clean	3,55	2,74	2,61
10 dB Hoth	PCM 8bit	2,75	2,35	2,36
10 dB Hoth	GSM 06.10	1,94	1,48	1,33
10 dB Hoth	MELPe 2.4	2,16	1,91	1,85

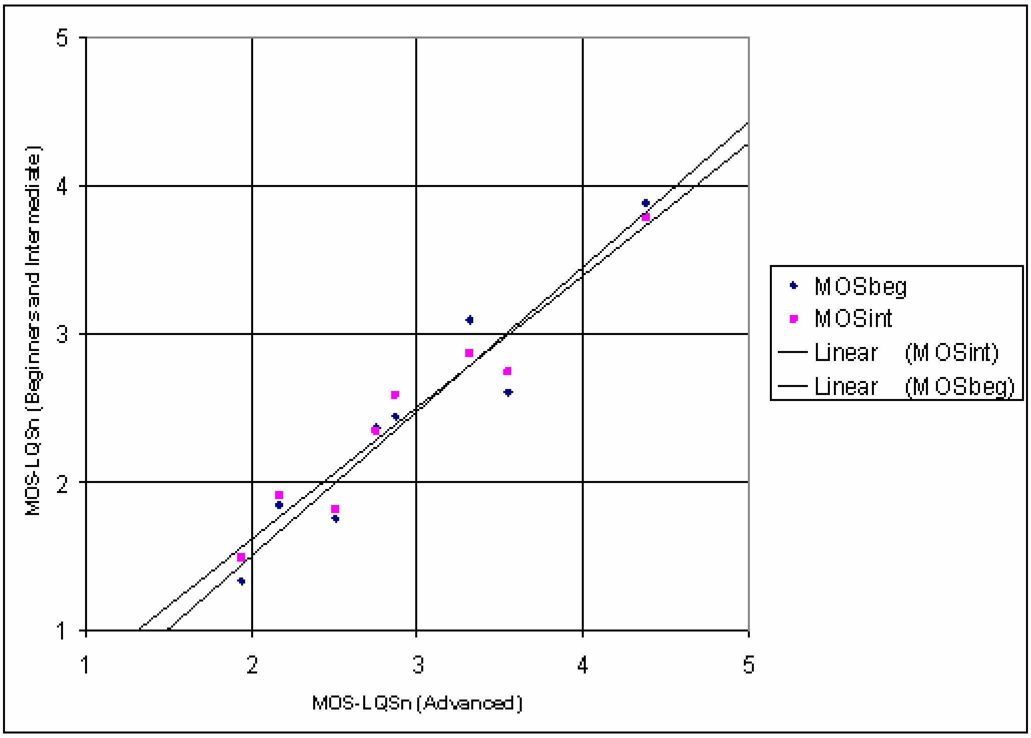

Fig.1 Subjective test results per condition, comparison between Advanced and both other groups

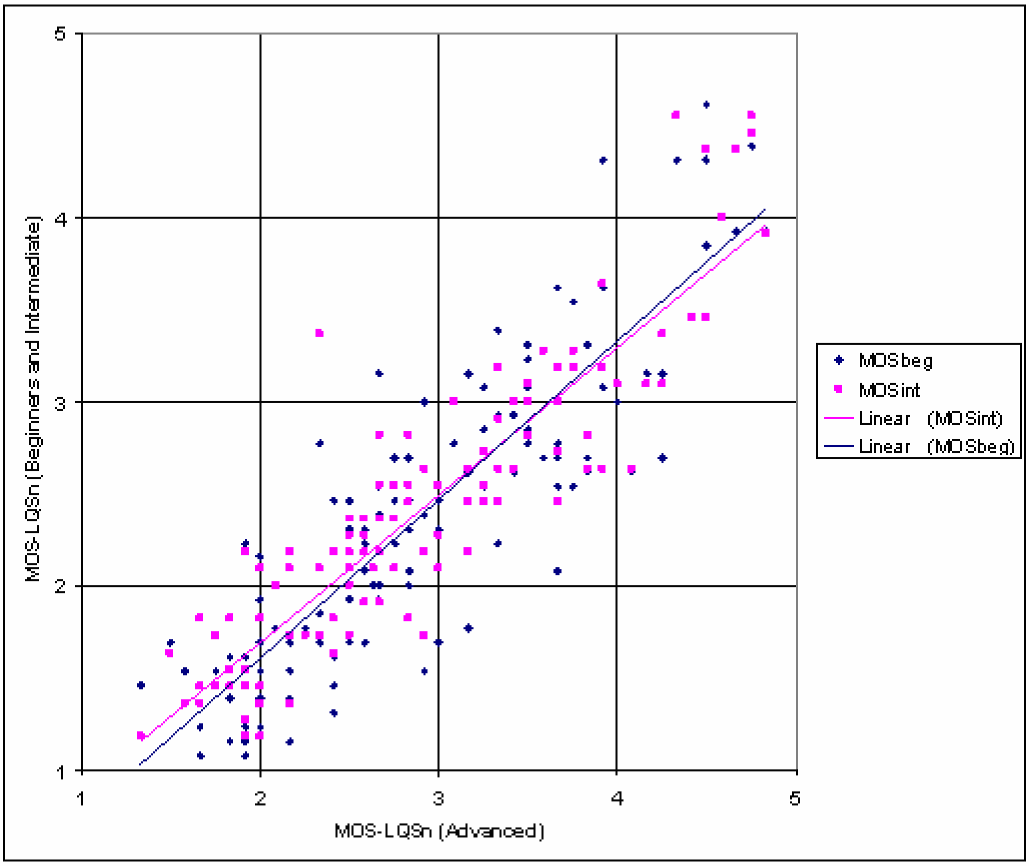

Fig. 2 Subjective tets results per sample, comparison between Advanced and both other groups