Assessment of Different Loudness Models for Perceived Speech Quality

Nicolas Côté[1,2], Valérie Gautier-Turbin[1], Sebastian Möller[2], Alexander Raake[2]
[1] *France Télécom R&D, 22300 Lannion, France, Email: nicolas.cote@orange-ftgroup.com*
[2] *Deutsche Telekom Laboratories, TU-Berlin, 10587 Berlin, Deutschland*

Introduction

The loudness is one of the perceived features involved in the assessment of the overall quality of transmitted speech. Impact of loudness on speech quality was shown to be bandwidth-dependent [1]. Attenuation of the speech signal was one of the most problematic point in the first analogue telephone transmissions. Nowadays, thanks to the packetized transmission, this should not be a problem anymore. However, speech enhancement algorithms involved in networks (e.g. mobile networks) may introduce time-variable attenuation/amplification on the speech. For this reason, four loudness models were selected to assess loudness of speech signals. Then, the loudness impairment factor $I_{e,loud}$ is introduced. This factor was derived from results of a speech quality auditory experiment. Estimated parameters from our four loudness models are compared to auditory $I_{e,loud}$ and then used to model this impairment factor.

Loudness Models

Modeling of the loudness perception is still under study, however several models are in-use for decades. They are separate in two groups: single-band models, and multi-band models. The models from the first group are based on time-domain information which is integrated into a one-dimensional parameter describing the energy of the signal. The models in the last group are mostly based on the theory of critical bands and level compression. After a calibration phase, both types of models reduced the information into a one-dimensional parameter.

Single-band model

The first selected model, called Equivalent Sound Pressure Level (Leq), is a measure of the mean energy of the signal. Several filters, as the well-known A filter, are applied on the measure to be closer to the perceived loudness of the signal. An algorithm using this Leq value combined with a filter was recently standardised as the ITU-R Recommendation BS.1770. It estimates the perceived loudness of audio sample. In addition, the ITU-T has defined the Active Speech Level (ASL) as the Leq of the speech-only parts of the sample [2], using a Voice Activity Detection (VAD).

Fletcher Model

The third model is the one from Fletcher who described in [3] a first multi-band model in order to compute the frequency-dependent loss L_{ME} introduced by a speech transmission channel. This model is used in order to estimate the energy which is necessary to render the input speech transmitted by a circuit (like a telephone path) perceptively equal in loudness to the distorted output speech. Based on Fletcher's study, a certain amount of level loss called 'Loudness Ratings' (LR) should introduce a certain amount of degradation of the overall quality. In our study wideband LR are estimated based on coefficients from Annex G of ITU-T Rec. P.79. However, this measure does not take into account time-varying loudness perception effect. Nowadays, a standardized instrumental method used to quantify speech quality, the so-called 'E-model' [4], is based on these Loudness Ratings.

Zwicker Model

The fourth model is the Zwicker model [5] which is a standardized multi-band method for determining loudness levels from an instrumental measure. However, the Zwicker's model was developped at first for steady sounds, such as tones or noise bursts and was then improved for temporally variable sounds. This method is used in speech quality models as a starting point for the perceptual representation of the speech stimuli.

Loudness Impairment Factor $I_{e,loud}$

Our aim is to quantify the degradation for speech listened at a non-optimum level with a specific loudness impairment factor $I_{e,loud}$ which should be compatible with the E-model. The optimum level corresponds to the speech level which gives the highest auditory quality score. However, the actual speech level used in auditory experiments corresponds to the prefered speech level[1] which is some dB lower than the optimum speech level. In addition, the difference between optimum and preferred level is dependent on other features such as the bandwidth or the signal-to-noise ratio. Auditory $I_{e,loud}$ were derived from a speech quality experiment presented in [1]. This experiment was carried out using an ACR 5-point scale. Degradations were simulated and then normalized to several desired levels by means of the Active Speech Level (ASL) tool [2]. Carrying out an auditory test is expensive and time-consuming. Because of this, an algorithm which estimate $I_{e,loud}$ from the speech signals were developped based on the selected loudness models. A methodology has been developed in [6] for deriving impairment factors from the results of auditory listening-only tests. This method was applied in order to derive $I_{e,loud}$. The procedure consist in transforming the MOS values to the an overall quality scale at the basis of the E-model (see [4]); on this scale, degradations are assumed to reduce the

[1] This preferred level was set to 79 dB_{SPL} for a monaural telephone-band listening situation.

maximum quality score of 129 (for the optimum wideband speech transmission at the preferred speech level), e.g. by a value of $I_{e,loud}$ for non-optimum loudness.

Estimation

Table 1: Correlation coefficient between the auditory MOS and the loudness parameters

ASL	Leq	Fletcher	Zwicker
0.963	0.969	-0.964	0.878

As a first step, the four loudness models described above were applied to speech stimuli used in the auditory test. Table 1 shows the Pearson correlations between the auditory MOS values of condition G.722.2@23.85 kb/s played at seven speech levels and the loudness parameters. We see quite a high correlation for all the parameters except for the Zwicker model. Estimations from this model do not have a linear relationship with the auditory MOS values. Table 2 shows the Pearson correlations and prediction errors between the auditory and estimated MOS values. We analyse two speech quality models, the parametric E-model [4] and intrusive PESQ [7]. For the first one, in a previous study [1] quality values were estimated based on fixed LR values for each condition. Based on Fletcher model, LR values are estimated in this study for each speech stimuli. Table 2 shows a slightly higher correlation and a lower prediction error using estimated LR values instead of fixed ones. For the latter speech quality model, in [1] we described the difficulties of the PESQ model to estimate speech quality of signals (reference and degraded) having different levels. Using the Zwicker model, modifications are proposed in order to improve the PESQ model. First the $I_{e,loud}$ can be modeled by:

$$I_{e,loud} = 0.433 * (\exp(\delta_{loud}/6.318) - 1) \qquad (1)$$

where δ_{Loud} corresponds to the loudness difference in sones between the reference speech file normalized at the optimum speech level and the degraded speech file. Then, estimated $I_{e,loud}$ are subtracted from the PESQ estimation mapped to the overall quality scale. Table 2 shows a higher correlation and lower prediction error for the corresponding enhanced PESQ compared to PESQ. Figure 1 shows the relationship between the auditory and estimated MOS values for the E-model (based on estimated LR values) and the enhanced PESQ.

Table 2: Correlation coefficient and prediction error between the auditory and estimated MOS values

Model	ρ	σ
PESQ	0.814	0.548
E-model	0.908	0.711
Enhanced PESQ	0.920	0.401
E-model (Fletcher)	0.913	0.526

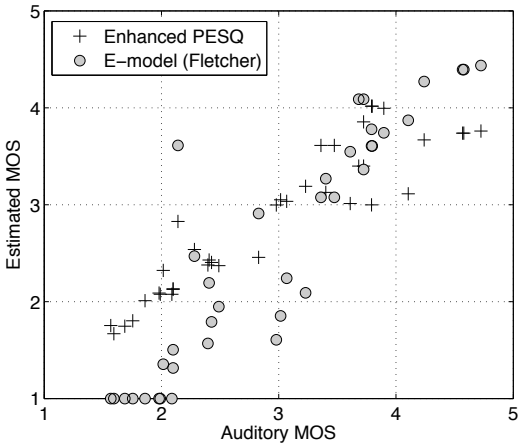

Figure 1: Relationship between the auditory and estimated MOS values

Conclusions

In this study, four loudness models were selected in order to improve the instrumental assessment of perceived overall quality in case of signals presented at a non-optimum level. Parameters estimated from loudness models are compared to auditory speech quality scores and then used to model loudness impairment factors. The introduction of a loudness measure in speech quality models as the parametric E-model or the signal-based PESQ model, seems to improve the estimations of such models.

References

[1] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of loudness level on the overall quality of transmitted speech," in *Proc. 123rd Conv. of the Aud. Eng. Soc.*, NY US, Oct. 5–8 2007.

[2] ITU–T Rec. P.56, *Objective Measurement of Active Speech Level*, International Telecommunication Union, CH–Geneva, 1993.

[3] H. Fletcher and R. H. Galt, "The perception of speech and its relation to telephony," *J. Acous. Soc. of America*, vol. 22, no. 2, pp. 89–151, 1950.

[4] ITU–T Rec. G.107, *The E-Model, a Computational Model for Use in Transmission Planning*, International Telecommunication Union, CH–Geneva, 2005.

[5] E. Zwicker, *Psychoakustik*, Springer Verlag, Berlin, 1982.

[6] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment factor framework for wideband speech codecs," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 6, pp. 1969–1976, 2006.

[7] ITU–T Rec. P.862, *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, CH–Geneva, 2001.