An Intrusive Super-Wideband Speech Quality Model: DIAL

Nicolas Côté[1], Vincent Koehl[1], Valérie Gautier-Turbin[2], Alexander Raake[3], Sebastian Möller[3]

[1]LISyC EA 3883, UBO/ENIB, Brest, France
[2]France Télécom R&D, Lannion, France
[3]Deutsche Telekom Laboratories, TU Berlin, Germany
cote@enib.fr

Abstract

The intrusive speech quality model standardized by the ITU–T shows some limits in its quality predictions, especially in a wideband transmission context. They are mainly caused by strong differences in perceived quality when speech is transmitted over different telephone networks. Instrumental methods should provide reliable estimations of the integral speech quality over the entire perceptual speech quality space. This paper presents a new model, called Diagnostic Instrumental Assessment of Listening quality (DIAL). It combines a core model, four dimension estimators and a cognitive model, providing integral quality estimations as well as diagnostic information in a super-wideband context.

Index Terms: Speech quality, super-wideband, estimation, intrusive measurement

1. Speech Quality

Following the point of view of Jekosch [1], quality is the result of the judgement of the perceived composition of an entity with respect to its desired composition. In the specific case of *speech quality*, the entity corresponds to an acoustic speech signal. Auditory tests are the most reliable way to assess the perceived speech quality. In such methods, subjects are asked to judge the quality of transmitted or otherwise processed speech signals. According to Raake [2] and based on ideas developed by Jekosch [1], the speech quality judgement process can be decomposed on a time scale in four successive steps (see Fig. 1):

1 Perception

The acoustic speech signal is perceived by the listener and results in a *perceived auditory composition*. The auditory composition includes all perceptual aspects such as the phonetic information and the characteristics of the talker and of the listener's environment. Such heterogeneous information, which are not yet related to quality, imply a multidimensional organization of all the perceptual aspects. It results directly that a listener can distinguish two acoustic speech signals on the basis of their perceived aspects. Several characteristics of the listener, such as his motivation, memory, knowledge, experience, and expectations influence the perception process. In addition to these personal characteristics, the context (i.e. the listener's environment) in which the sound occurs also contributes to the perception process and therefore to the speech quality. Both types of characteristics form the *response modifying factors* which adjust the *desired auditory composition* to a particular listening situation.

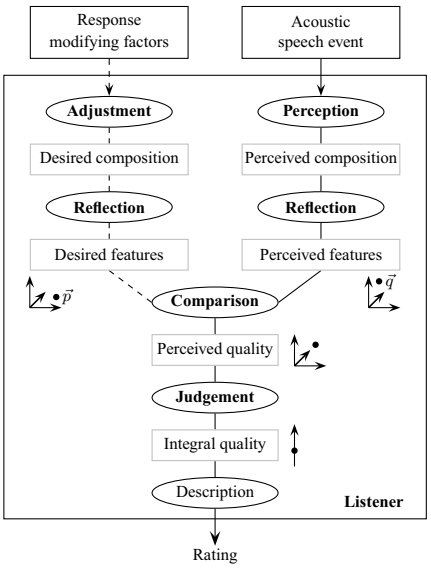

Figure 1: *Speech quality judgement process as seen from the listener's perspective, according to Raake [2] and based on Jekosch [1].*

2 Reflection

The listener reflects on all the signal characteristics which are relevant for quality, i.e. "names" each feature of the multidimensional space[1]. These "nameable" features are related to quality. The perceived composition is thus defined by a set of values (one per feature), i.e. the perceived features, describing a position $\vec{q}$ in the multidimensional perceptual space. In parallel, the same "nameable" features are used to define a position $\vec{p}$ for the desired composition, i.e. the desired features.

3 Comparison

The quantification of the quality of the acoustic speech signal requires a comparison of the desired and perceived features $\vec{q}$ and $\vec{p}$, i.e. their corresponding values for each "nameable" feature.

4 Judgement

The listener judges the quality using comparison. The judgement process corresponds to an aggregation of all

[1]It is worth noting that the features are not literally named by the listener. On the contrary, the reflection is a "subconscious" process.

features into a single quality value, i.e introducing a weighting coefficient to each feature related to its influence on the quality. The acoustic speech sample is thus of high quality only if the listener's perception is identical to the desired composition (i.e. similar values for the perceived and desired "nameable" features).

To achieve high quality, good comprehensibility and intelligibility of the transmitted speech are a necessary prerequisites. A consensus exists between authors: speech quality, like other perceptual magnitudes, is considered as a "multidimensional" object: the user's comparison process can be represented in his brain by a few orthogonal "nameable" features, called *perceptual dimensions* (of speech quality) [1]. Following the taxonomy used by Möller [3], the term *integral quality* is used when the user's judgement encloses all these perceptual dimensions.

2. Instrumental methods

Since auditory tests are costly and time-consuming, instrumental methods have been developed. These methods, referred to as *quality models*, correspond to a computer program designed to automatically estimate the perceived quality of speech signals thanks to a simulation of the human perception. Such a simulation is based on a mathematical model which provides a relationship between a sensation and a physical magnitude.

Since 2001, the "Perceptual Evaluation of Speech Quality" (PESQ) is the most widely used instrumental model, see ITU–T Rec. P.862 [4]. The PESQ model estimates the perceived quality of transmitted speech for the classical Narrow-Band (NB, 300–3400 Hz) telephone bandwidth. In 2005, the PESQ model was extended to the evaluation of WideBand (WB, 50–7000 Hz) transmissions. This WB mode of PESQ, called WB-PESQ, has been standardized as the ITU–T Rec. P.862.2 [5]. The algorithm of WB-PESQ is very similar to the one used by PESQ for NB signals. However, WB-PESQ uses exclusively speech signals with a sampling frequency of $f_S = 16$ kHz. Still, WB-PESQ shows several limitations. For instance, it has difficulties to reliably predict the perceived quality of WB low bit-rate speech codecs [6]. Some modifications of the perceptual model of WB-PESQ have been introduced in [6] to improve WB-PESQ reliability. They result in a modified version of the WB-PESQ model, called Mod. WB-PESQ. Other studies have focused on the time-alignment algorithm of PESQ: WB-PESQ sometimes under-estimates the quality of packet-switched networks introducing time-warping effects [7].

Nowadays, Voice over IP (VoIP) applications enable Super-WideBand (S-WB, 50–14000 Hz) transmissions. However, WB-PESQ is not able to assess the quality of the corresponding S-WB speech codecs, e.g. ITU–T Rec. G.722.1 Annex C [8]. Since no instrumental model has been developed for this bandwidth yet, a new model is required in order to assess the quality of all in-use speech processing and transmission systems.

3. DIAL

The following section describes a new intrusive model, called "Diagnostic Instrumental Assessment of Listening quality" (DIAL) which has been developed as part of the ITU–T standardization program called "Objective Listening Quality Assessment" (P.OLQA). The P.OLQA project aims at standardizing a new intrusive speech quality model. Intrusive models use a reference (clean or system input) speech signal $x(t)$ and a corresponding degraded (distorted or system output) speech signal $y(t)$. Speech quality models usually estimate the integral quality of the degraded speech files. DIAL follows the assumption that the combination of several specialized measures is more efficient than one single complex measure. This model relies on a specific framework which combines three building blocks:

A core model
: It estimates the non-linear degradations introduced mainly by speech processing systems such as low bit-rate codecs.

Four dimension estimators
: They quantify the degradations on the four perceptual dimensions: *Directness/Frequency Content*, *Noisiness*, *Continuity* and *Loudness*.

A cognitive model
: An aggregation of all the estimated scores simulates cognitive processes employed by the human listener during the quality judgement process.

3.1. Core model

The *Core* model is based on the "Telecommunication Objective Speech-Quality Assessment" (TOSQA) model [9]. The TOSQA model was modified to cover non-linear degradations. Such degradations are estimated by a comparison of the perceptually transformed reference $x(t)$ and degraded $y(t)$ signals. The perceptual transformation[2], resulting in $L_x(l,z)$ and $L_y(l,z)$, are mainly based on the model of loudness calculation developed by Zwicker and Fastl [10]. Then, the distortions which are perceptually irrelevant (i.e. not audible) are compensated before the comparison of $L_x(l,z)$ with $L_y(l,z)$ by the *Core* model. The perceptual comparison corresponds to a similarity measure. The final *Core* model quality score MOS_{Core} is defined on the Mean Opinion Score (MOS) scale.

3.2. Dimension estimators

Three of the four dimension estimators cover the perceptual quality space derived by Wältermann [11].

1. *Directness/Frequency Content (DFC)*
2. *Noisiness*
3. *Continuity*

However, several studies (e.g. McDermott [12]) introduced the listening level as an additional feature of the integral speech quality. Consequently, an estimator for the perceptual dimension *Loudness* has been included as well. Each estimator quantifies the perceived quality on one of these four perceptual dimensions. The resulting quality-score framework is referred to as "degradation decomposition".

The estimator for the quality dimension *DFC* measures the linear frequency degradation introduced by a transmission system. The *DFC* estimator uses a perceptual representation of the frequency response of the system. Two parameters are estimated using the method developed by Scholz [13], the bandwidth in terms of an Equivalent Rectangular Bandwidth (ERB) and the center frequency (f_c) of the frequency response in Hz. These two parameters are combined according

[2]Here, $l = 1 \ldots L$ corresponds to the frame (time) index (16 ms length) and $z = 1 \ldots 24$ corresponds to the Bark scale (frequency) index

to the model developed by Raake [2] providing a bandwidth impairment factor I_{bw}. The I_{bw} is then mapped to the MOS scale resulting in a MOS_{DFC} value.

The second estimator for the dimension *Noisiness* combines different algorithms which have been developed especially for the DIAL model. The first algorithm estimates the additive noise in the degraded signal $y(t)$ using the "silence/noisy" (i.e. without speech) frames only. The estimated parameter corresponds to a noise loudness value L_n. A discontinuous transmission (DTX) algorithm will avoid the transmission of the signal in silence/noisy frames. In this case, the environmental noise at the talker's side is transmitted during speech periods only, resulting in an under-estimated additive noise loudness value L_n. A "Noise on Speech" (NoS) parameter quantifies the additive noise components during speech periods only. The final *Noisiness* score MOS_N is calculated using the maximum degradation value estimated by the two parameters L_n and NoS.

A third estimator for the perceptual dimension *Continuity* has been developed by Huo [14]. This estimator uses Weighted Spectral Slope (WSS) distances and signal temporal loss to derive three parameters (i) the interruption rate r_I, (ii) the artefact rate r_A, and (iii) the clipping rate r_C. The MOS_C value is then calculated using a non-linear combination of these three parameters.

The *Loudness* estimator quantifies the degradation for speech heard at a non-optimum listening level. An optimum level corresponds to the speech level which leads to the highest auditory quality score. This estimator calculates an equivalent sound pressure level L_{eq} which corresponds to the mean energy of $y(t)$ over all speech frames. Then, a *Loudness* quality MOS_L value is derived from the parameter L_{eq}.

3.3. Cognitive model

Many intrusive quality models simulate the human peripheral auditory system (i.e. represent the signal at the output of the inner ear). According to Beerends [15], this paradigm shows limitations. In order to increase the accuracy of instrumental measures, a model of the cognitive processes should be included. Ideally, the instrumental measures should interpret the perceptual dimensions involved in the assessment process like a human listener would do. Therefore, DIAL includes a cognitive model which determines the influence of each perceptual dimension on the integral quality.

Cognitive processes are generally modelled by machine learning techniques. For instance, Pourmand et al. [16] used a Bayesian modelling to estimate the quality of noise reduction algorithms. The DIAL model provides an integral speech transmission quality estimation MOS_{LQO} and four additional quality estimates, MOS_{DFC}, MOS_{L}, MOS_{C} and MOS_{N} according to the four perceptual dimensions. These perceptual estimators described in the previous section are used to diagnose the quality degradations. The DIAL instrumental measure has been obtained by combining the dimension estimators to the *Core* model to form a reliable quality speech quality model.

In DIAL, the combination of the estimated quality features with the *Core* model estimation relies on a machine learning technique, called k-Nearest Neighbours (kNN). This is a non-parametric method used for different purposes such as density estimation, identification or instrumental assessment. The algorithm uses two data sets $\{\vec{x}_1 \dots \vec{x}_N\}$ and $\{y_1 \dots y_N\}$ comprising $n = 1, \dots, N$ observations (i.e. speech stimuli). Each observation n includes a vector $\vec{x}_n = (x_{n,1}, \dots, x_{n,5})$ of five estimated quality values and a corresponding auditory integral quality score y_n. The parameters i correspond to the MOS estimations from the four perceptual estimators and the *Core* model.

Considering a test vector $\vec{\mu} = (\mu_1, \dots, \mu_5)$ (an unknown stimulus), the goal of the kNN algorithm is to assign an integral quality score MOS_{LQO} to $\vec{\mu}$. This algorithm comprises four stages:

1. The parameter values $\vec{x}$ and $\vec{\mu}$ are normalized to the range $[-1, +1]$.
2. The Euclidean distance to the test value $\vec{\mu}$ for all N observations is calculated as:

$$D_n = \sqrt{\sum_{i=1}^{5} (\mu_i - x_{n,i})^2} \tag{1}$$

   where n is the observation index and i the parameter index.
3. The K observations having the lowest distance are selected as the "neighbours" of the test value $\vec{\mu}$.
4. The estimated integral quality score corresponds to the averaged integral quality scores over the selected K neighbours:

$$MOS_{\mathrm{LQO}} = \frac{1}{K} \sum_{k=1}^{K} y_k \tag{2}$$

3.4. Evaluation of DIAL

This evaluation corresponds to a comparison of the DIAL estimations to auditory quality judgements. This paper presents only the performance of the integral speech quality estimations MOS_{LQO}. Overall, DIAL has been compared to 39 WB and S-WB auditory experiments. To reliably evaluate the newly developed model, DIAL is compared to the reference intrusive speech quality measures including the WB-PESQ [5], the Modified WB-PESQ [6], the TOSQA [9] and the "PErception MOdel–Quality assessment" (PEMO–Q) model developed by Huber and Kollmeier [17]. Table 1 shows the overall Pearson correlation coefficients (ρ) and Root Mean Squared Errors (RMSE, σ) over the 39 databases. The DIAL model obtains the best correlation $\rho_{\mathrm{DIAL}} = 0.913$ and lowest RMSE $\sigma_{\mathrm{DIAL}} = 0.380$.

Figure 2 shows the "per-condition" DIAL estimations for an example S-WB database. This database has been carried out at France Télécom R&D (Lannion, France) in accordance with ITU-T Rec. P.800 [18]. A 5-point ACR listening quality scale has been used. The test corpus includes conditions impaired by linear degradations such as non-optimum listening levels, background noises, discontinuities, bandwidth restrictions and a combination of them. It reflects thus the whole speech quality space. The DIAL model obtains a correlation of $\rho_{\mathrm{DIAL}} = 0.80$ and a RMSE of $\sigma_{\mathrm{DIAL}} = 0.60$. This relatively low correlation is obtained since few conditions are under-estimated. For instance, conditions 12 and 13 are strongly under-estimated.

Table 1: *Experimental results over* 39 *WB and SWB auditory tests of five intrusive speech quality models including DIAL. These measures are computed after third order mapping function.*

Ranking order	Model	Correlation ρ	RMSE σ
1	DIAL	0.913	0.380
2	Mod. WB-PESQ	0.879	0.443
3	WB-PESQ	0.856	0.480
4	TOSQA	0.826	0.524
5	PEMO-Q	0.749	0.616

These conditions have been re-scaled at a new sampling frequency of $f'_S = 1.1 \times f_S$ and $f'_S = 0.9 \times f_S$. Conditions 10 and 11, corresponding to digital overload played back at a normal listening level, are under-estimated by the *Core* model. These conditions introduce strong non-linear distortions. Conditions 19 and 35 are also under-estimated by DIAL. These conditions are impaired by real background noise (street and car noise respectively). However, neither the *Noisiness* estimator nor the *Core* model under-estimate these conditions. These problems may be caused by the cognitive model.

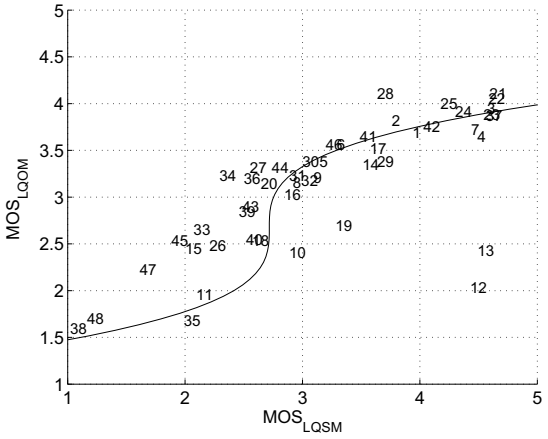

Figure 2: *DIAL estimations for an example S-WB Database. The curve represents the estimated third order mapping function.*

4. Conclusions

The new quality model is evaluated on a large set of databases. Several reference instrumental models including the current ITU–T standard (WB-PESQ) are compared to DIAL. Overall, DIAL outperforms all instrumental reference models, including WB-PESQ and the Modified WB-PESQ. However, the S-WB operational model of DIAL fails to predict the quality of some specific conditions. For instance, codec tandeming conditions are under-estimated by DIAL. Such inaccuracies come from either the *Core* and dimensions quality estimations or the cognitive model. Further developments are needed to improve the DIAL model and to obtain reliable estimations of the integral quality over the whole speech quality space.

5. References

[1] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. DE–Berlin: Springer, 2005.

[2] A. Raake, *Speech Quality of VoIP - Assessment and Prediction*. UK–Chichester: Wiley, 2006.

[3] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. USA–Boston, MA: Kluwer Academic Publ., 2000.

[4] *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 Std. ITU–T Rec. P.862, 2001.

[5] *Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs*, International Telecommunication Union Std. ITU–T Rec. P.862.2, 2005.

[6] N. Côté, V. Gautier-Turbin, A. Raake, and S. Möller, "Analysis of a Quality Prediction Model for Wideband Speech Quality, the WB-PESQ," in *Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, DE–Berlin, September 2006, pp. 115–122.

[7] N. Shiran and I. Shallom, "Enhanced PESQ Algorithm for Objective Assessment of Speech Quality at a Continuous Varying Delay," in *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, 2009, pp. 157–162.

[8] *14 kHz Mode at 24, 32, and 48 kbit/s*, International Telecommunication Union Std. ITU–T Rec. G.722.1 Annex C, 2005.

[9] J. Berger, *TOSQA Telecommunication Objective Speech Quality Assessment*, ITU–T Del. Contrib. COM 12–34, CH–Geneva, 1997.

[10] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*, 1st ed. DE–Berlin: Springer, 1990.

[11] M. Wältermann, K. Scholz, A. Raake, U. Heute, and S. Möller, "Underlying Quality Dimensions of Modern Telephone Connections," in *Proc. 9th International Conference on Spoken Language Processing (ICSLP 2006)*, USA–Pittsburgh, PA, September 17–21 2006, pp. 2170–2173.

[12] B. J. McDermott, "Multidimensional Analyses of Circuit Quality Judgments," *Journal of the Acoustical Society of America*, vol. 45, no. 3, pp. 774–781, 1969.

[13] K. Scholz, M. Wältermann, L. Huo, A. Raake, S. Möller, and U. Heute, "Estimation of the Quality Dimension "Directness/Frequency Content" for the Instrumental Assessment of Speech Quality," in *Proc. 9th International Conference on Spoken Language Processing (ICSLP)*, USA–Pittsburgh, PA, September 17-21 2006, pp. 1523–1526.

[14] L. Huo, M. Wältermann, U. Heute, and S. Möller, "Estimation of the Speech Quality Dimension "Discontinuity"," in *Proc. 8th ITG-Fachbericht-Sprachkommunikation*, DE–Aachen, October 8–10 2008.

[15] J. Beerends, "Modelling Cognitive Effects that Play a Role in the Perception of Speech Quality," in *Proc. Workshop on Speech Quality Assessment*, DE–Bochum, November 10–11 1994, pp. 2–9.

[16] N. Pourmand, D. Suelzle, V. Parsa, Y. Hu, and P. Loizou, "On the Use of Bayesian Modeling for Predicting Noise Reduction Performance," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, TW, April 19–24 2009, pp. 3873–3876.

[17] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[18] *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union Std. ITU–T Rec. P.800, 1996.