Attribute-based Instrumental Assessment for Speech- Transmission Quality

Lu Huo, Ulrich Heute
University of Kiel, Germany, Email: {lhu,uh}@tf.uni-kiel.de

Introduction

Although, as state-of-the-art, PESQ [1] has already proved its excellent performance in predicting speech transmission quality both in the industry and in the laboratories, it can deliver only a single MOS (Mean-Opinion-Score) prediction and does not provide more insight of the underlying causes of the perceived quality degradation. An attribute-based method, as an alternative to integral-quality models such as PESQ, assumes that the overall speech quality is based on a weighted sum of different perceptual dimensions/attributes in the human hearing, and thus the objective model could be built in two steps:

1. Predict each quality dimension/attribute $\hat{D}_i$ based on objective measurements;
2. predict the integral speech quality $\widehat{MOS}$ by a weighted superposition of the estimated perceptual quality dimensions:

$$\widehat{MOS} = b_0 + \sum_{i=1}^{L} b_i \cdot \hat{D}_i. \qquad (1)$$

In a previous paper by M. Waeltermann and K. Scholz et. al. [2], a framework and initial realization of an attribute-based instrumental measure for end-to-end speech transmission quality was given. Three dimensions, namely, *discontinuity* (DIS), *noisiness* (NOI) and *coloration* (COL) had been identified through multi-dimensional analysis, instrumental measures had been developed to measure each of these dimensions, which are indexed by so called dimension impairment factors, and a prediction model was also given.

This paper summarizes and enhances the further development published in [3, 4, 5]. The following modifications are included: (1) Pre-processing is refined and efforts have been taken to separate different attributes so that their mutual influences are suppressed; (2) the extracted parameters and the prediction model of each dimension are modified, based on the implication of the so-called "sub-dimensions" (SD) identified by follow-up auditory tests and multidimensional analysis.

Fig. 1 demonstrates the overall structure of the proposed model. As we can see, the whole processing can be subdivided into three blocks: Preprocessing, dimension estimator, and integral model, which are described respectively in the following sections.

Preprocessing

The purpose of the preprocessing is to separate the influence of degradations regarding to the identified dimensions and to provide intermediate representations of signals or distortions for the dimension estimators. Via resampling, we work with a sampling frequency of 16 kHz.

1. Prior to any other operation, a simple time alignment is carried out to compensate the constant delay between the clean signal $x(k)$ and the degraded signal $y(k)$ because no jitter exists in our database. A noise floor of -70 dBov is added to both signals and the resulting signals are adjusted to -26 dBov active signal level after ITU-T Rec. P. 56.
2. To provide information for the COL estimator, the linear distortion is firstly estimated in the form of a system frequency response $H(\Omega)$ by comparing the spectra of the clean and degraded signals. Here a long Hamming window of length 8192 (i.e., 512 ms) and 50% overlap is used. This length is used to include also long impulse responses as those in the case of hands-free terminals (HFT).
3. Pre-distortion is then applied to the original signal, in order to compensate the influence of linear distortion in measuring the other two dimensions. But this is not a trivial task. Due to the presence of strong non-linear distortions, such as background noise, or packet-loss, the exact measurement of the system's linear distortion can sometimes be impossible, especially since test stimuli are often only 2 to 3 seconds long. The resulting error can cause large inaccuracy in the later estimations. Thus, 5 alternative pre-filter designs, based on $H(\Omega)$ or its smoothed versions [5], are tried. The one that causes the smallest average weighted spectral slope (WSS) [6] distance between the pre-filtered original spectrum and the degraded-signal spectrum is chosen. After pre-filtering, the short-time spectra are obtained by fast-Fourier transformation (FFT) using a short Hamming window of length 256 (16ms) and 75% overlap.
4. Auditory critical-band filters are applied to both the clean and the degraded spectra for each frame, resulting in critical-band spectra. Here, 34 critical bands are used for 0-8 kHz.
5. To provide the information for the DIS estimator, the WSS metric and the energies for both degraded and original signal are obtaind on a frame basis. The

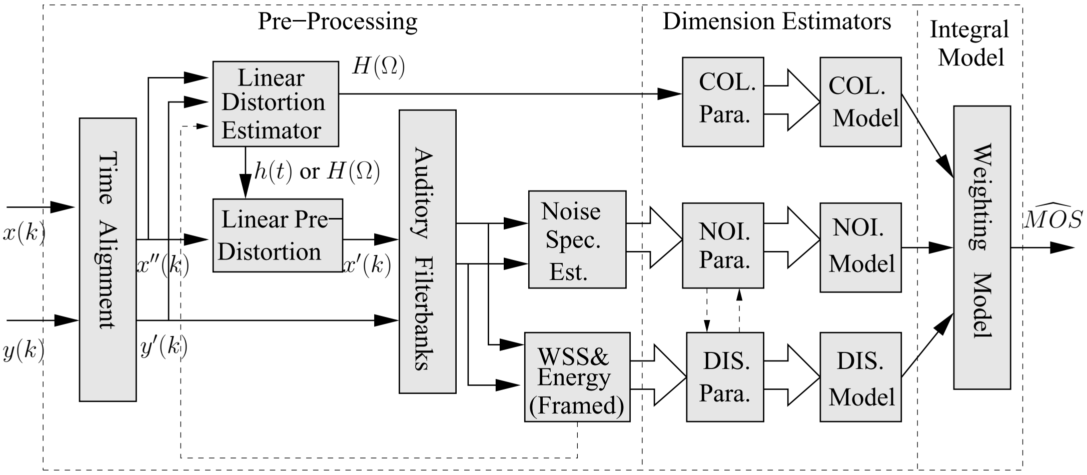

Figure 1: Overall structure of the proposed instrumental attribute-based speech quality assessment

WSS metric was originally developed to estimate the vowel intelligibility degradation [6], and was also found useful to identify strong non-linear distortions in the spectrum [5]. The average WSS also serves as a selection criterion for the pre-distortion filter.

6. To provide the information for the NOI estimator, the noise spectrum is firstly obtained as the difference between the critical-band spectra of the degraded and original signals during speech pauses and then weighted by the A-curve according to [7]. Additionally, the cepstral distance is measured between the original and degraded spectrum.

Dimension Estimators

The internal representations of signals or distortions provided by the preprocessing block are further processed in three threads, in each of which dimension scores are predicted based on the extracted dimension parameters. Interactions take place between the NOI and the DIS estimators, as indicated in Fig 1: NOI should provide a raw frame-wise signal-noise-ratio (SNR) for calculating the thresholds for packet-loss detection in DIS, and DIS should provide the positions of packet-loss-influenced frames so that their influence on NOI can be suppressed.

In order to study the dimensions and train the dimension estimators, specific databases for each dimension have been collected, thanks to our project partner in D-Telekom Laboratories (T-Labs), Berlin. For each database, a multidimensional-scaling (MDS) experiment was designed and carried out to reveal the perceptual space, and sub-dimensions (SDs) have been found. Based on these three databases, dimension estimators have been developed in [3, 4, 5]. Table 1 summarizes these databases, the identified SDs, the parameters responsible for each SD, and the prediction performances. For the details of the databases and SD identification, the readers are referred to the above mentioned publications.

Dimension 1: Coloration

Coloration is mainly associated with linear distortions. Two SDs, namely, *directness* and *frequency content*, have been identified [3] for this dimension. Three parameters are found responsible:

- *Equivalent rectangular bandwidth ERB* and *gravity center of frequency* θ_G: These two parameters represent the basic shape of the estimated gain function. To obtain ERB and θ_G, the estimated system frequency response should firstly be expressed on a dB scale and transformed from a linear Hz scale to the Bark scale, in order to simulate the loudness perception of human hearing. Then, the warped gain function is shifted by ST so that the main part of the curve lies above zero. ERB is calculated as the area between the curve and zero, divided by the maximum of the curve; θ_G is calculated as the gravity center of this area [8].
- *Reverberation time* T_{30}: To highlight the influence of the room acoustics introduced by HFT, 30 dB reverberation time T_{30} is calculated as the time of the 30 dB decay of the system impulse response.

Dimension 2: Discontinuity

Discontinuity is mainly associated with strong nonlinear distortions such as packet-loss, packet-loss concealment (PLC), time-clipping, or musical noise. MDS analysis has revealed three to four SDs for this dimension: *interruptedness*, *additive artifacts*, *musical noise classification*, and an extra SD for time-clipping.

Four parameters are found to be the suitable:

- *Interruption rate* r_I as the percentage of frames where interruptions, either artifacts or energy dips, take place. Two types of interrupts have been detected. The first one, called “energy dips”, happens when lost segments are muted. This type can be relatively easily detected by sudden disappearance and reappearance of the degraded-signal energy or large energy loss compared to the original-signal energy. The second type, called “additive artifacts”, happens when lost segments are substituted by PLC but the distortions can be still perceived. This type of interrupts cannot be detected simply from the energy criterion; thus the

Dimension	Database conditions	SDs	Responsible features
COL (80×2)	Real and simulated recording in rooms and cars (7), conventional and mobile handsets (12) simulated frequency response(61)	Directness Frequency Content	Equivalent bandwidth and 30-dB reverberation time Gravity center of frequency
DIS (74×2)	G.711, G.728, G.729A and codec-free × PL% of 0, 3, 5, 10, 15, and 20 (58), front-end-clipping (3) simulated musical noise (13)	Interruptedness Additive Artifacts Musical Noise Clipping rate	Interruption rate Artifact rate Musical noise level Clipping rate
NOI (69×2)	Additive (57): circuit noise, "subscriber-line" noise and ambient noise Multiplicative (12): ADPCM noise, MNRU	Speech Contamination Additive Noise Level Noise Coloration	Speech-band noise level Additive noise level Bright noise level

Table 1: *Resulting Quality-dimensions from Multidimensional Scaling by T-Labs. The first column contains the abbreviations of each attribute-specific database as well as the number of conditions × speakers, the second column shows the included conditions and its number in bracket for each database, the third column indicates the identified sub-dimensions, and the last column summarizes the responsible parameters measured for each sub-dimension.*

weighted spectral slope (WSS), as a metric for the spectrum distortion, is applied: When a sudden large spectrum distortion happens without accompanying energy drop, the corresponding frames are marked as "additive artifacts".

- *Artifact rate* r_A as the percentage of frames with artifacts. It was observed that a frame with audible "additive artifacts" is more annoying than a simple muted frame. Thus the artifact rate constitutes a single SD by itself.
- *Clipping rate* r_C as the percentage of frames with time-clipping. Time-clipping can be explained as a large section of muted frames. In our case, time clipping is identified when more than 0.1 sec speech is detected as muted.
- *Musical noise level* n_{mn}. Musical noise is a well-known phenomenon when a noise-reduction algorithm is used in the transmission system and the residual noise possesses a random, tone-like characteristic. In our study, musical noise is detected in speech pauses using a so-called "relative approach" [9], i.e., the musical noise is separated from the comparatively wide and stationary noise floor by detecting noise peaks in a certain searching range of frequency-time blocks in the spectrogram.

It is worth mentioning that till now it is still open whether the additive artifacts and the musical noise belong exclusively to the dimension noisiness, and it is very likely that these two kinds of degradations influence both dimensions. However, in our study, they are temporally treated under the dimension discontinuity only.

Dimension 3: Noisiness

Noisiness, as the name suggests, summarizes the perceptual influence of all kinds of noise on the speech quality.

Studies in [4] show that this dimension can be further decomposed into three SDs: *speech contamination*, *additive noise level*, and *noise coloration.* Thus the following parameters are extracted:

- *Speech-band noise level* n_{lf}. This first SD is correlated with the noise-energy level of the speech-relevant spectral band actually during the speech activity, be it an additive noise or a signal-correlated noise. Two alternative methods are used to represent this measure: In the case that strong noise energy detected in speech pauses is above a threshold, its level is averaged over speech-band frequencies; otherwise, the weighted cepstral distance is measured to catch the noise that probably exists only during speech activities.
- *Additive noise level* n_p is measured by the noise level averaged over the whole frequency range in pauses.
- *Bright noise level* n_{hf}. The last SD singles out strongly colored noise types such as realistic background noise and the very "bright" subscriber-line noise. This parameter is measured as the high-frequency noise level in cases when the gravity center of noise is above a threshold (10 Bark in our case).

Prediction Models and Performances

Sub-dimension-based prediction models of the global dimensions are firstly trained by the attribute-specific databases listed in Table 1. Readers are referred to [3, 4, 5] for details. The MOS values of each database serve as the dimension scores of the corresponding dimension scores. Special care has been taken so that the resulting prediction should be highly correlated with the MOS value of the corresponding database and at the same time be uninfluenced by the other dimension predictions.

Table 2 shows the correlations between the prediction results and the MOS values of the three databases, which is an indication of the good prediction performance and the good orthogonality of the proposed dimension estimators.

	$\widehat{D1}$	$\widehat{D2}$	$\widehat{D3}$
$D1$	0.93	0.33	0.26
$D2$	0.28	0.93	0.17
$D3$	-0.29	0.39	0.94

Table 2: *Performance and orthogonality of the proposed dimension estimators. $D1$, $D2$ and $D3$ refer to the MOS values collected for each attribute-specific database, $\widehat{D1}$, $\widehat{D2}$ and $\widehat{D3}$ to the predicted dimension scores for each database.*

Then, the overall prediction model is trained by a mixed database Q. This database was collected by T-Labs during a common research project; it is composed of all possible types of single distortions, selected from the attribute-specific databases, and some combinations of distortions crossing the dimensions. The description of this database can be found in [2]. From four speakers available in this database, the authors used one of them as the training data to establish the relation of the dimension scores and the overall speech quality, and used the rest as the test data. The simple linear regression strictly following (1) is employed here although a higher-order prediction model involving interaction terms is also highly probable, but to avoid over-training it is postponed in the future research. To demonstrate the strengths and weaknesses of the proposed model, the well-known public databases in ITU-T P. supplement 23 (only test databases exp1 and exp2 are available to the authors) are used as test databases and the performance of the state-of-the-art algorithm PESQ is displayed for comparison.

Table 3 shows the results. For the database Q, the proposed model seems to outperform PESQ. This may be simply due to the fact that our model is trained by these conditions or has already "seen" the conditions. However, it may also result from the fact that the proposed model takes into account the different influences from different types of distortions and is more similar to the human judging process of the speech quality, and thus can handle complicated distortion scenarios better than yielding only integral single distortion scores. Nevertheless, in the case of the "unseen" database ITU-T P.sup. 23, PESQ delivers a far superior prediction performance. This, on the one hand, shows the need to further improve the proposed model in the case of pure codec distortion, which dominates these two databases, and, on the other hand, shows the potential of further improvement: After all, the simplest prediction model is used in our case and it can already deliver a sound results. By experiment, simple retraining the overall prediction model using the P.sup. 23 databases can improve the correlation to above 0.9. Thus more efforts are needed in the training phase of the integral model.

Test	Cond. × Speaker	proposed	PESQ
Database Q (seen database)			
Training	69× 1	0.87	0.77
Test	69× 3	0.86	0.77
ITU-T P.sup. 23 (unseen database)			
Exp1	44 × 12	0.82	> 0.93
Exp3	50 × 16	0.81	> 0.93

Table 3: *Performance of the proposed MOS predictor, comparing with PESQ. The first column names the tests, the second column shows the different conditions and speakers available in each test, the third and the last column show the correlations of the prediction results and the target MOS values by the proposed method and PESQ, respectively.*

Summary and Outlook

In this paper, an attribute-based prediction model is introduced, which, besides a MOS prediction, also provides the highly desirable scores of specific attributes as diagnostic information. Although it can up to now only deliver a moderate prediction performance ($\rho \approx 0.81 \sim 0.86$) through its simple form, it can be improved by more sophisticated prediction model.

In the future, more efforts will be put into, firstly, establishing a stable and advanced prediction model, secondly, establishing better measures for the fine distortions as those from codecs, and thirdly extending the prediction model to the wideband situation.

Acknowledgements

The authors would like to thank the "Deutsche Forschungsgemeinschaft" (DFG) for the financial support of the present work under the grants HE 4465 and MO 1038 and also S. Möller's group at T-Labs for their excellent co-operation.

References

[1] ITU-T Rec. P.862: *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-end Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs.* ITU-T, Geneva, 2001.

[2] Wältermann, M., Scholz, K., Möller, S., Huo, L., Raake, A., and Heute, U.,: *An Instrumental Measure for End-to-end Speech Transmission Quality Based on Perceptual Dimensions: Framework and Realization.* Interspeech 2008, Brisbane, Australia, 2008.

[3] Huo, L., Wältermann M., Heute, U. and Möller, S.: *Estimation Model for the Speech-Quality Dimension "Directness / Frequency Content".* IEEE Workshop on Applications of Signal Processing to Audio Acoustics (WASPAA2007), New York, 2007.

[4] Huo, L., Wältermann M., Heute, U. and Möller, S.: *Estimation Model for the Speech-Quality Dimension "Noisiness".* 5th European Congress on Acoustics (Forum Acusticum 2008), Paris, France, 2008.

[5] Huo, L., Wältermann M., Heute, U. and Möller, S.: *Estimation of the Speech Quality Dimension "Continuity".* 8th ITG Conference on Speech Communication, Aachen, Germany, 2008.

[6] Klatt, D.: *Prediction of perceived phonetic distance from critical-band spectra: a first step.* IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, 1982.

[7] ITU-R Rec. BS. 468: *Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting.* ITU-T, Geneva, 1990.

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

[9] Genuit, K.: *Objective Evaluation of Acoustics Quality Based on a Relative Approach.* 25th International congress on Noise Control Engineering, Liverpool, 1996.