3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

Diagnostic Instrumental Speech Quality Assessment in a Super-Wideband Context

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

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

Abstract

Speech quality models usually estimate the integral quality of the degraded speech files. Such quality values do not inform system developers and telephone service providers on the perceived degradation introduced by the system under study. This paper describes a new intrusive speech quality model, called Diagnostic Instrumental Assessment of Listening quality (DIAL), providing diagnostic information in both narrow-band and super-wideband contexts. Contrary to previous methods, this model estimates scores on four perceptual quality dimensions, *Directness/Frequency content*, *Continuity*, *Noisiness* and *Loudness*. These four dimensions are assumed to define the whole speech quality space.
Index Terms: speech quality, diagnostic, super-wideband, intrusive measurement

1. Introduction

Speech processing system developers and telephone service providers are interested in the Quality of Experience (QoE) of prototype and in-use speech transmission systems and speech processing systems. In order to quantify the QoE, auditory tests, such as those described in ITU–T Rec. P.800 [1], are the most reliable measurement methods. They assess the perceived quality of processed or otherwise transmitted speech signals. However, such tests are costly and time-consuming. Therefore, instrumental measurement methods have been developed. Among the different types of instrumental methods, intrusive signal-based models such as PESQ in ITU–T Rec. P.862 [2] provide highly reliable quality estimations. Intrusive models use a reference (clean or system input) speech signal $x(k)$ and a corresponding degraded (distorted or system output) speech signal $y(k)$. Across 22 different sets of experimental data, an average correlation coefficient of $\rho = 0.935$ was obtained between the PESQ estimations and the auditory quality scores.

The PESQ model is an integral model, i.e. quality estimations are reflected by a single integral quality score. With this single quality value, however, it is not possible to reveal the causes of the estimation: equal integral quality scores might be obtained for a condition impaired by a certain degree of audio bandwidth restriction on one hand, and a condition with a certain level of background noise on the other hand. Moreover, multiple different types of degradations occurring simultaneously might change in such a way that the integral quality remains the same. In such cases diagnostic measures are desirable which are capable of describing the underlying quality dimensions. These measures decompose the integral quality into several attributes.

The perceived quality of speech signals and its multidimensional character are defined in Section 2. Then, drawbacks and advantages of several diagnostic models are detailed in Section 3. A new diagnostic model, called DIAL, based on specific framework, is described in Section 4. This new model is evaluated in Section 5.

2. Speech Quality

Following the point of view of Jekosch [3], 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 such as those described in ITU–T Rec. P.800 [1] are the most reliable way to assess the perceived speech quality. In such methods, subjects are asked to rate the quality of speech signals. According to Raake [4] and based on ideas developed by Jekosch [3], the speech quality rating process can be decomposed on a time scale in five successive steps (see Figure 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, linguistic knowledge and telecommunication experience 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.

2 Reflection
The listener reflects on all the signal characteristics which are relevant for quality, i.e. "names" each feature of the

10.21437/PQS.2010-12

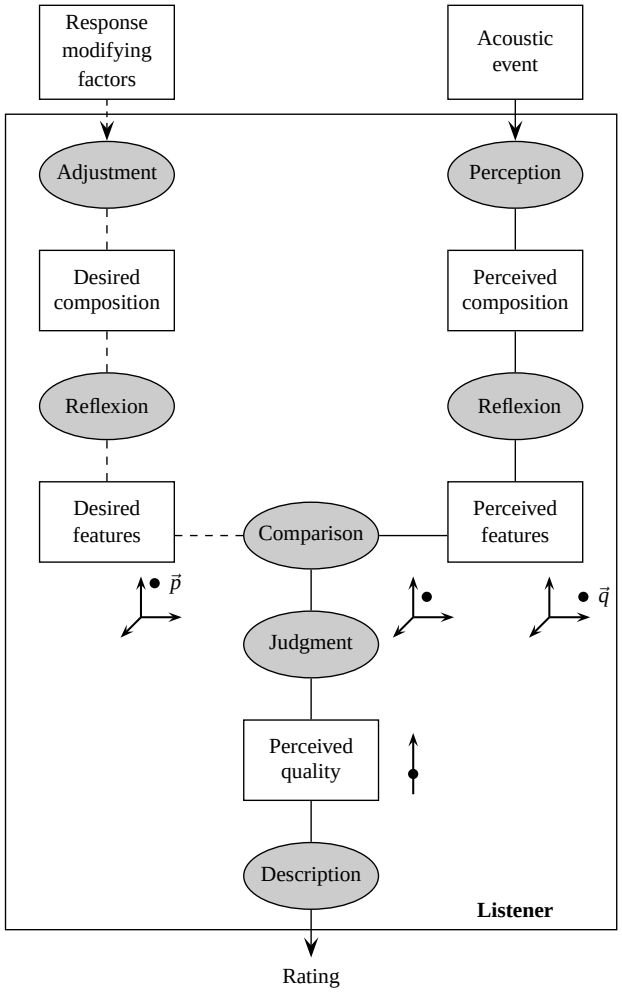

Figure 1: *Speech quality rating process as seen from the listener's perspective, according to Raake [4] and based on Jekosch [3]*

multidimensional space. These "nameable" features, called *perceptual dimensions* [3], are related to quality and orthogonal in the speech quality space. 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 features into a single quality value, the *integral quality* [5], i.e introducing a weighting coefficient to each feature related to its influence on the quality. The acoustic speech sample is thus of quality only if the listener's perception is identical to the desired composition (i.e. similar values for the perceived and desired "nameable" features), or even exceeds the desired composition.

5 Description
The listener finds the best possible description of the perceived quality using the rating scale. In case the scale corresponds to the 5-point listening quality scale defined in ITU–T Rec. P.800 [1], the listener chooses one of the five categories *bad*, *poor*, *fair*, *good* or *excellent*.

3. Diagnostic Models

Auditory test methods defined in ITU–T Rec. P.800 [1] and integral models such as PESQ [2] provide a single quality score. A *diagnostic* method decomposes the integral quality into several characteristics. For instance, the system under study can be characterized by several physical attributes such as its overall gain, its frequency response and its Signal-to-Noise Ratio (SNR). However, these information are useless for the end user and do not inform about the influence of system parameters involved in the user's perception. Ideally, these characteristics should correspond to dimension values, according to the phase "Comparison" in the rating process. Jekosch in [3] gives a definition of a diagnostic measure. A **diagnosis** is

> "[the] production of a system performance profile with respect to some taxonomization of the space of possible inputs."

Quackenbush et al. in [6] developed a set of four quality feature estimators corresponding to scales used in the auditory test method called Diagnostic Acceptability Measure (DAM) [7]. Halka and Heute in [8] and Moore et al. in [9] developed signal-based models for quantifying two speech quality features (i) the *linear* degradations and (ii) the *non-linear* degradations. Another approach has been followed by Gierlich et al. in [10]. The authors developed a diagnostic model called Relative Approach (RA) for the diagnostic assessment of Narrow-Band (NB, $f \in [300, 3400]$ Hz) and WideBand (WB, $f \in [50, 7000]$ Hz) communications in the presence of background noise. This diagnostic model estimates two quality features (i) the *speech* signal quality and (ii) the *background noise* quality. Beerends et al. in [11] developed a diagnostic model which is mainly based on the intrusive quality model PESQ [2]. This model estimates the following three MOS quality feature values: *noise* degradations, *frequency* degradations and *time-varying* degradations.

However, such diagnostic measure should rely on orthogonal perceptual dimensions defined in Section 2. These dimensions can be derived from a multidimensional analysis of the auditory results. Sen in [12] developed a new set of five dimension estimators. The corresponding perceptual dimensions have been selected using a Principal Component Analysis (PCA) analysis applied to DAM auditory results.

4. DIAL

Recently, Wältermann et al. in [13] derived a speech quality space using two independent auditory test methods (i) a paired-comparison similarity test and (ii) a Semantic Differential (SD) test. Using a MultiDimensional Scaling (MDS) analysis (applied to the similarity test results) and a PCA (applied to the SD

results), a stable speech quality space has been derived. This space is defined by the following three perceptual dimensions:

- *Directness / Frequency content* (*DFC*)
- *Continuity*
- *Noisiness*

However, several studies (e.g. McDermott [14]) 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. This set of four perceptual dimensions is assumed to cover the whole speech quality space including modern telephone networks. Therefore, the new diagnostic model described in this paper relies on these four dimensions.

4.1. Overview

The following section describes a new intrusive diagnostic 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. 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 (see Figure 2) 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.

Dimension estimators
They quantify the degradations on four perceptual dimensions.

A cognitive model
An aggregation of all the quality feature estimations into an integral quality score simulates cognitive processes employed by the human listener during the quality judgement process.

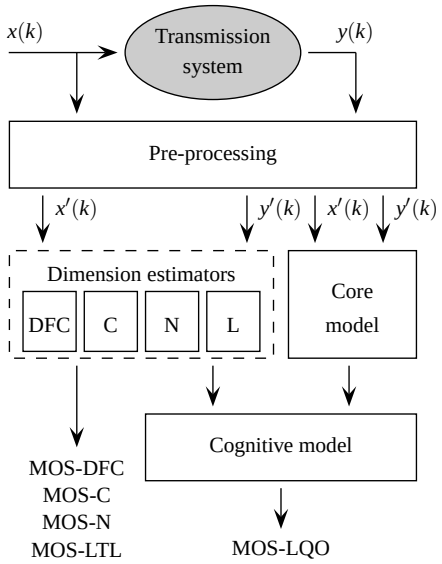

Figure 2: *Overview of the DIAL model*

The DIAL model provides a quality-score framework composed of 5 MOS values. The cognitive model gives an integral quality score (MOS-LQO) and each dimension estimator quantifies the perceived quality on one of the four perceptual dimensions (MOS-DFC, MOS-C, MOS-N and MOS-LTL). In addition, DIAL has two operational modes, a NB mode and a Super-WideBand (S-WB, $f \in [50, 14000]$ Hz) mode.

This paper focuses on the set of four dimension estimations. This set is composed of a *DFC* estimator developed by Scholz and Heute [15] and a *Continuity* estimator developed by Huo [16], both extended to S-WB transmissions. Two new estimators have been developed for the dimensions *Noisiness* and *Loudness*. The next paragraphs describe these four dimension estimators.

4.2. Directness/Frequency Content

The perceptual dimension *Directness/Frequency Content* (*DFC*) is related to the characteristics of the frequency response of the overall transmission system (i.e. mouth-to-ear), see [13]. The quality feature *Frequency content* covers the impact of bandwidth restrictions on the degraded speech file $y(k)$. *Directness* includes specific impairments such as the influence of the talking-room reflections and the effect of "coloration" (i.e. *dark* or *bright*) introduced by the transducers in a user terminal (e.g. headsets and telephone handsets).

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 $H(e^{j\Omega})$. Then, two parameters are estimated from $H(e^{j\Omega})$ using an algorithm developed by Scholz et al. in [15]: (i) the bandwidth in terms of an Equivalent Rectangular Bandwidth (*ERB*, in Bark), and (ii) the Central frequency (f_c, in Hz) of the frequency response. The original algorithm has been extended to cover S-WB conditions.

The two estimated parameters are combined according to the model developed by Raake [4] providing a bandwidth impairment factor I_{bw}. The I_{bw} is then mapped to the MOS scale, using the transformation described in ITU–T Rec. G.107 [17], resulting in a MOS-DFC value.

4.3. Continuity

Continuity degradations correspond to either an isolated distortion or a time-varying distortion. An isolated distortion is caused by the loss of one or several speech frames, by erroneous bits during radio transmissions or by time-clipping introduced by Voice Activity Detection (VAD) algorithms. In the worst case, the lost or discarded frames are replaced by silence frames of the same length, called "zero insertion". However, a Packet Loss Concealment (PLC) algorithm can reduce the impairment of such isolated distortions using an interpolation from the previous and/or next frame.

The detection of "discontinuities" is highly influenced by the right alignment of $x(k)$ and $y(k)$. Therefore, the preprocessing stage of DIAL uses the robust time-alignment algorithm used in PESQ [2] which has been extended to cover time-warping conditions, i.e. including continuous variable delay. The estimator for the dimension *Continuity* has been developed by [16]. This estimator detects the discontinuities in the speech

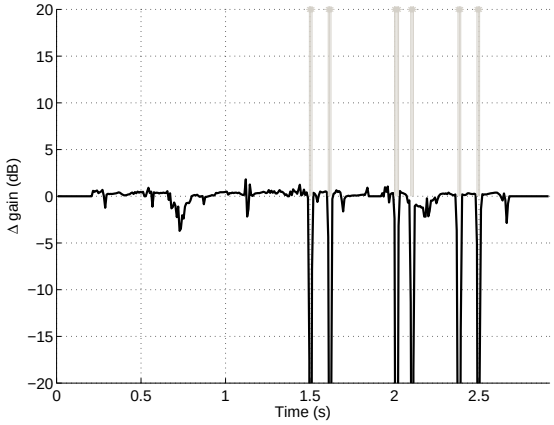

Figure 3: *System gain variation (**black curve**) and detected zero insertions (**vertical gray lines**) for an example condition.*

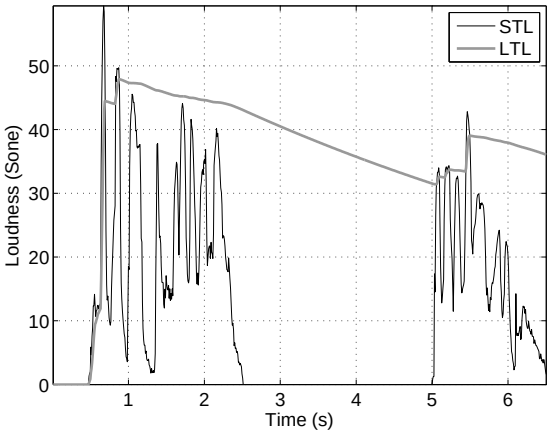

Figure 4: *Short-term loudness and corresponding temporal integration for an example condition.*

signal using Weighted Spectral Slope (WSS) distances [18] and the system gain variation (Δ gain). Then, it derives three parameters:

- Short level variation rate r_L, e.g. zero insertion.
- The artifact rate r_A, e.g. audible spectrum deviations in interpolated frames.
- Interruption rate r_I, e.g. time-clipping.

The MOS-C value is calculated using a non-linear combination of these three parameters. This estimator has been developed for a Wideband (WB) context and it has been slightly modified to be applied in a S-WB context. Figure 3 shows the system gain variation (Δ gain) and detected zero insertions for an example condition.

4.4. Noisiness

Different types of noise may impact differently the perceptual dimension *Noisiness* such as (i) environmental noises at the talker's side, (ii) circuit noises introduced by analog transmissions, and (iii) coding noises introduced by waveform coders, e.g. ITU–T Rec. G.726 [19].

The *Noisiness* estimator combines different algorithms which have been developed especially for the DIAL model. The first algorithm estimates the additive noise level in the degraded signal $y(t)$ using the "silence/noisy" (i.e. without speech) frames only. However, discontinuities such as interferences in transducers produced by mobile transmissions may result in an over-estimated additive noise level. Therefore, this algorithm includes a detection of discontinuities in silence/noisy frames. The resulting parameter is a noise loudness value (L_n, in Sone).

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*, in dB) 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*.

4.5. Loudness

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 integral quality score. This algorithm estimates the perceived loudness of the whole speech signal. Firstly, a Short-Term Loudness (STL, in Sone) is calculated for each active speech frame using the loudness model for stationary sounds developed by Zwicker in [20]. Then, the STL values are aggregated over the time scale into a Long-Term Loudness (LTL, in Sone) according to a model developed by Glasberg and Moore in [21]. Figure 4 shows the short-term loudness values and the temporal integration for an example condition. The LTL parameter corresponds to the integrated value at the last active frame. The LTL parameter is then mapped to the MOS scale into a *Loudness* score MOS-L.

5. Evaluation of DIAL

In this section, DIAL estimations are compared to auditory test results unknown during the development process. For this purpose, a WB auditory test carried out following the methodology described in [22] is used. In this auditory test method, the subjects are asked to rate the speech stimuli on an ACR 5-point listening quality scale and on 3 continuous dimension scales *DFC*, *Noisiness* and *Continuity*. The test corpus includes, background noises, packet-loss conditions, several WB and NB low bit-rate speech codecs and codec tandeming conditions. Figure 5 shows the relationship between DIAL diagnostic estimations and the auditory scores. A third order mapping function has been applied on DIAL estimations for each dimension. This mapping function reduces the influence of the test corpus on the evaluation process. The DIAL model is used in the S-WB operational mode. The "Per-condition" Pearson correlation coefficients ρ and prediction errors σ between the auditory and estimated MOS scores are presented in Table 1. These statistical measures are calculated using the auditory dimension scores and the scores provided by the corresponding estimators.

DIAL estimators provide relatively good quality dimension

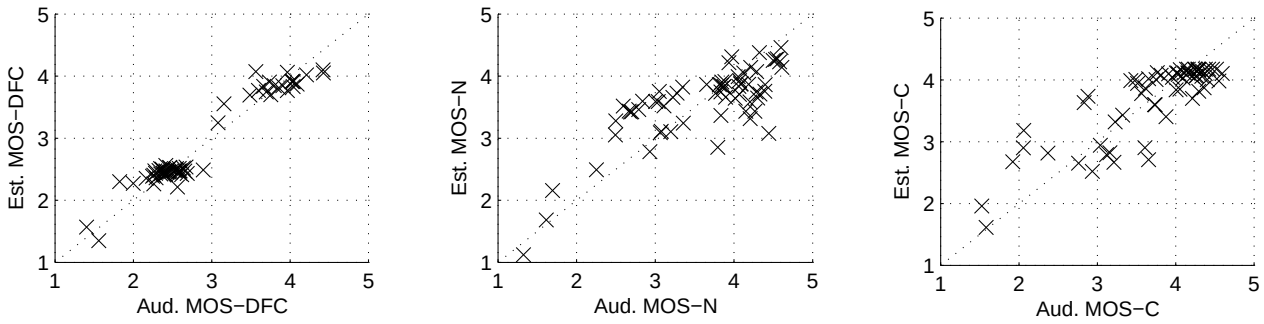

Figure 5: *Example degradation decomposition for a WB auditory test. This figure shows the estimated and auditory quality scores for the* 3 *dimensions* DFC, Noisiness *and* Continuity.

estimations, especially for the Dimension *DFC*. Deviations to the auditory scores are observed for Dimension *Noisiness*. This auditory test does not follow the usual input signals framework. The averaged percentage of vocal activity of the input speech signals is close to 80% (instead of 50% in other auditory tests). This prevents a reliable estimation of the additive noise level parameter L_n. Here, the *Noisiness* dimension is estimated by the parameter "Noise on Speech" (*NoS*). For the Dimension *Continuity*, DIAL estimator is less reliable than for the *DFC* dimension. This estimator under-estimates codec tandeming conditions.

Dimension estimations can be used to "diagnose" the degradation introduced by each condition under test. Figure 6 presents the relationship between DIAL diagnostic estimations and the integral auditory MOS values. The test corpus includes different types of background noise, packet losses, bandwidth restrictions and low bit-rate speech codecs. For this database, only the integral auditory quality scores are available.

For the *DFC* dimension, Figure 6 shows the three bandwidths, NB, WB and S-WB. These three bandwidths are well delimited excepts for some S-WB conditions which are under-estimated. The lowest one corresponding to a condition which have been re-sampled with a different sampling frequency, i.e. time re-scaling. This effect impacts also the *Continuity* estimator. For the *Noisiness* dimension, some noisy conditions are over-estimated. These conditions include other degradations such as packet losses which may result in an under-estimated additive noise level L_n. For the *Continuity* dimension, the re-sampled condition is largely under-estimated and some other non degraded conditions (on this dimension) are slightly under-estimated. The continuity estimator detects discontinuities in case of codec tandeming conditions. In addition, a wrong alignment of the two speech signals results in under-estimations. For the *Loudness* dimension, an amplification implies an increase of the estimated MOS-L values. This effect simulates that the optimum level (highest MOS-L values) is higher than the preferred listening level, i.e. the default level in auditory tests [1].

Table 1: *Pearson correlation coefficients* ρ *and prediction errors* σ *for the* DFC, Continuity *and* Noisiness *dimension estimators.*

Dimension	ρ	σ
DFC	0.968	0.320
Continuity	0.846	0.397
Noisiness	0.783	0.486

6. Conclusions

A new diagnostic speech quality model, called DIAL, has been developed. This model has been evaluated on different unknown auditory tests. DIAL is the first intrusive model providing diagnostic information over the whole speech quality space and in a S-WB context. However, DIAL fails to reliably quantify specific conditions. For instance, time re-scaling and codec tandeming conditions are under-estimated by the *Continuity* estimator. Further developments are needed to improve the DIAL model and to obtain reliable estimations of all in use transmission systems.

7. References

[1] *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union Recommendation P.800, 1996.

[2] *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 Recommendation P.862, 2001.

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

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

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

[6] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. USA–Englewood Cliffs, NJ: Prentice Hall, 1988.

[7] W. D. Voiers, "Diagnostic Acceptability Measure for Speech Communication Systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'77)*, Hartford, 1977, pp. 204–207.

[8] U. Halka and U. Heute, "A New Approach to Objective Quality-Measures Based on Attribute-Matching," *Speech Communication*, vol. 11, no. 1, pp. 15–30, 1992.

[9] B. C. J. Moore, C. T. Tan, N. Zacharov, and V. V. Mattila, "Measuring and Predicting the Perceived Quality of Music and Speech

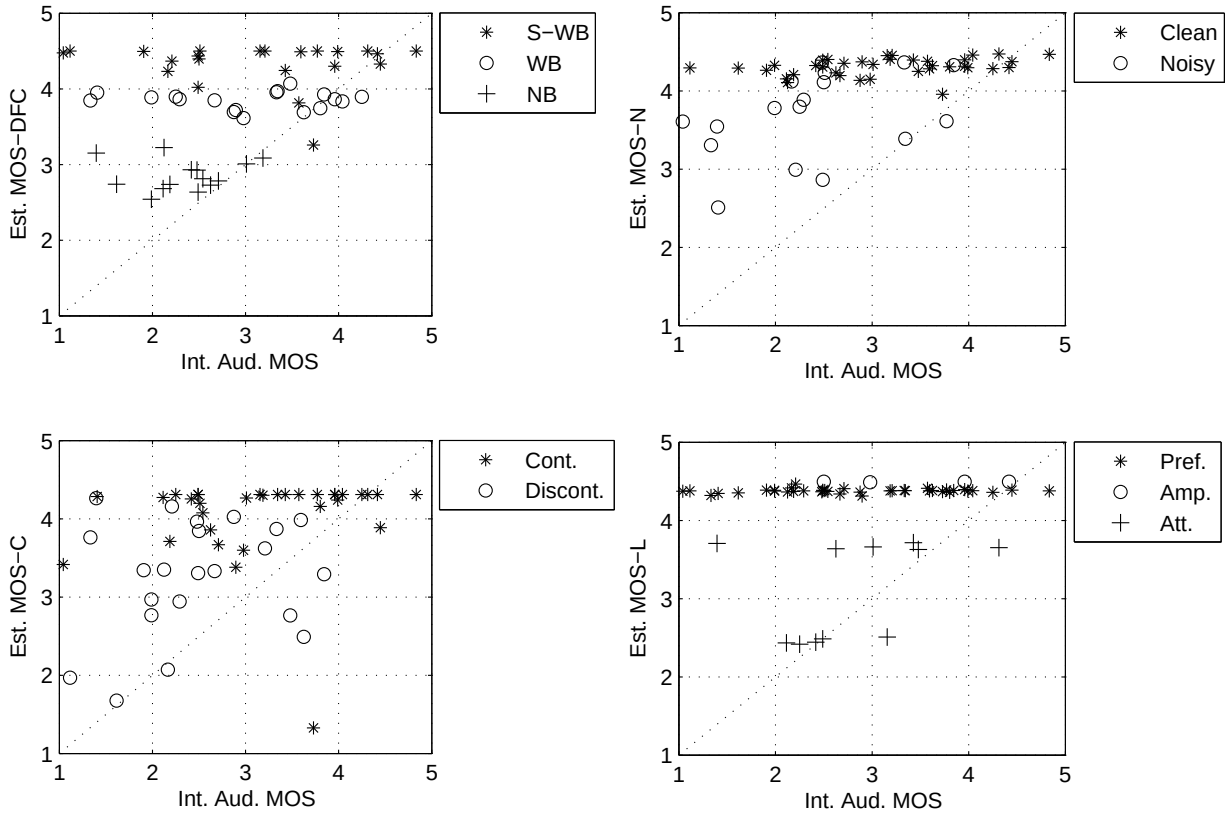

Figure 6: *Diagnostic information for an example S-WB Database. Abbreviations: 'Cont.' ≡ Continuous; 'Discont.' ≡ Discontinuous; 'Pref' ≡ Preferred listening level; 'Amp.' ≡ Amplification; 'Att.' ≡ Attenuation.*

Subjected to Combined Linear and Nonlinear Distortion," *Journal of the Audio Engineering Society*, vol. 52, no. 12, pp. 1228–1244, 2004.

[10] H. W. Gierlich, F. Kettler, S. Poschen, and J. Reimes, "A New Objective Model for Wide-and Narrowband Speech Quality Prediction in Communications Including Background Noise," in *Proc. 16th European Signal Processing Conference (EUSIPCO)*, CH–Lausanne, 2008.

[11] J. G. Beerends, B. Busz, P. Oudshoorn, J. Van Vugt, K. Ahmed, and O. Niamut, "Degradation Decomposition of the Perceived Quality of Speech Signals on the Basis of a Perceptual Modeling Approach," *Journal of the Audio Engineering Society*, vol. 55, no. 12, pp. 1059–1074, 2007.

[12] D. Sen, "Predicting Foreground SH, SL and BNH DAM Scores for Multidimensional Objective Measure of Speech Quality," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'04)*, vol. 1, CA–Montreal, May 17–21 2004, pp. 493–496.

[13] M. Wältermann, K. Scholz, A. Raake, U. Heute, and S. Möller, "Underlying Quality Dimensions of Modern Telephone Connections," in *Proc. 9th Int. Conf. on Spoken Language Processing (ICSLP'06)*, USA–Pittsburgh, PA, September 17–21 2006, pp. 2170–2173.

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

[15] 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 Int. Conf. on Spoken Language Processing (ICSLP'06)*, USA–Pittsburgh, PA, September 17-21 2006, pp. 1523–1526.

[16] 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.

[17] *The E-Model, a Computational Model for Use in Transmission Planning*, International Telecommunication Union Recommendation G.107, 1998.

[18] D. Klatt, "Prediction of Perceived Phonetic Distance from Critical-band Spectra: A First Step," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'82)*, vol. 7, May 1982, pp. 1278–1281.

[19] *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)*, International Telecommunication Union Recommendation G.726, 1990.

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

[21] B. R. Glasberg and B. C. J. Moore, "A Model of Loudness Applicable to Time-Varying Sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[22] M. Wältermann, A. Raake, and S. Möller, *Assessment of Speech Quality Dimensions: Methodology, Experiments, Analysis*, International Telecommunication Union Contribution COM 12–82, 2009.