Assessment of the Speech-Quality Dimension “Noisiness” for the Instrumental Estimation and Analysis of Telephone-Band Speech Quality

Kirstin Scholz1, Christine Kühnel2, Marcel Wättermann2, Sebastian Möller2, Ulrich Heute4

1 Institute for Circuit and System Theory, Christian-Albrechts-University of Kiel, Germany
2 Deutsche Telekom Laboratories, Berlin Institute of Technology, Germany
3 ks@tf.uni-kiel.de

Abstract

The development of instrumental measures that do not only estimate the speech quality of modern telecommunication systems but also analyze it is a current issue in speech processing. Our work aims at an analytic quality measure for telephone-band speech that is based on the instrumental assessment of so-called quality dimensions. They describe different quality-relevant characteristics of speech signals and thus allow for a quality analysis. An overall-quality rating is obtained by a suitable combination of the dimension ratings. For telephone-band speech, three quality dimensions have been identified: “directness/frequency content”, “continuity”, and “noisiness”. This paper presents an estimator for “noisiness” that is based on three parameters: the level and the color of additive noise contained in a speech signal as well as its amount of signal-correlated noise. The dimension estimates show a correlation > 0.95 with the results of auditory tests. Index Terms: instrumental speech-quality assessment, speech-quality analysis, quality dimensions, noisiness

1. Introduction

During the transmission over modern telecommunication systems, speech signals are subject to various disturbances and deformations. In general, such distortions affect the quality of the transmitted speech signals perceived by the systems’ users.

As speech transmission quality is an important aspect for the acceptability of modern telecommunication systems [1], its assessment is of great interest to system developers and operators. Since quality is - generally speaking - the result of human perception and human judgment, valid speech-quality ratings can only be obtained from human listeners within auditory tests.

The most commonly used auditory test is the listening-only test yielding a rating for a system’s overall speech-quality. This rating is denoted as mean-opinion score (MOS) of a system.

Due to the fact that the auditory assessment of speech quality is time-consuming and expensive, instrumental speech-quality measures are needed. These measures determine quality ratings much more efficiently than auditory assessment methods, but they can only estimate speech quality. Nevertheless, reliable instrumental measures can substitute or complement auditory tests in areas where the use of auditory methods is not reasonable or even impossible. Instrumental methods are regarded as reliable in case their MOS estimates show a high correlation with the results of auditory tests.

Well-defined signal-based measures exist for the instrumental assessment of a system’s overall speech quality in terms of MOS values. These include for example PESQ (Perceptual Evaluation of Speech Quality) [2] and TOSQA (Telecommunication Objective Speech-Quality Assessment) [3]. However, such measures might provide the same MOS for two systems which human listeners perceive to sound differently. Hence, more differentiating analytic quality measures are needed that allow for a better distinction between the quality of various systems. In addition, such a quality analysis might provide starting points for improving a system’s speech quality. A possible approach is the instrumental assessment and analysis of speech quality on the basis of quality dimensions.

Our work aims at a corresponding measure for assessing the speech quality of modern telecommunication systems with a bandwidth limitation to a maximum frequency of 4 kHz [4], that is, the quality of narrowband speech-signals y(k) influenced by up-to-date speech codecs and distortions such as packet loss, additive background noise, noise-reduction methods and hands-free terminals. Three quality dimensions have been identified as relevant for the assessment of telephone-band speech quality: “directness/frequency content”, “continuity”, and “noisiness” (Section 2). This paper presents an instrumental measure for the dimension “noisiness”. Based on its characteristics, parameters are defined that describe dimension-relevant features of a speech signal (Section 3). These parameters provide the basis for the dimension estimator. Its structure is presented in Section 4, its performance is discussed in Section 5. Section 6 summarizes the paper and gives an outlook to our future work.

2. Dimension-based assessment of telephone-band speech quality

Quality dimensions describe distinct quality-relevant features of speech signals, e.g., the naturalness and the presence of noise. In case N dimensions together describe all relevant aspects of speech quality, the MOS of a system can be obtained by a suitable combination of the dimension ratings. Quality dimensions ideally describe speech quality with as few as possible independent characteristics and, thus, are regarded as uncorrelated.

Three dimensions have been identified as relevant for assessing the overall quality of narrowband signals y(k) that have been transmitted over modern telecommunication systems [5]:

1. “Directness/frequency content”: This dimension characterizes the quality-relevant influence of a transmission system’s frequency response on the signal y(k).
2. “Noisiness”: This dimension indicates the perceptual influences of noise-like disturbances on the quality of the signal y(k), such as background noise or circuit noise.
3. “Continuity”: This dimension is associated with disturbances of the form of the signal y(k) in the time domain, e.g., packet losses or short insertions like musical noise.
Expressing the overall quality of a telephone-band speech signal $g(k)$ in terms of a weighted linear combination of these three quality dimensions yields a model for a system’s MOS that covers about 90% of the total variance of the speech-quality judgments obtained in auditory tests [5]. To realize an equivalent instrumental quality measure, estimators for the above-mentioned quality dimensions are needed. An estimator for the dimension “directness/frequency content” was presented in [6]. In the following, an approach for quantifying “noisiness” is introduced.

3. Quality dimension “noisiness”

3.1. Auditory experiments

The characteristics of “noisiness” are determined based on the results of three auditory tests. Two multidimensional analyses, a semantic differential (SD) and a multidimensional scaling (MDS), have been carried out with the objective of identifying the quality dimensions of telephone-band speech. For this purpose 28 so-called narrowband stimuli (NB-stimuli) have been used. These stimuli have been obtained by processing two German sentences spoken by one female and one male speaker each with fourteen test conditions. The test conditions represent different configurations of modern telecommunication systems for the transmission of narrowband signals. They cover simple codecs, bandpass filters, handsfree terminals as well as distortions due to packet loss, background noise, circuit noise and noise-reduction algorithms [5].

A third experiment has been conducted on the basis of 84 so-called noisiness stimuli (N-stimuli). Its aim has been to analyze “noisiness” in more detail. The N-stimuli have been generated by processing two German sentences spoken by one female and one male speaker each with 42 test conditions which are assumed to influence a speech signal only in the dimension “noisiness”. These conditions cover different types and levels of sending-side background noise (both, babble, and hammer noise), various levels of circuit and terminal noise as well as different intensities of signal-correlated noise produced by speech codecs. During the auditory test, the participants rated the quality of the N-stimuli on the MOS scale. It is assumed that the obtained scores reflect the ratings of “noisiness” for the N-stimuli as these stimuli are only influenced in this dimension.

3.2. Characteristics of “noisiness”

The results of the three above-mentioned auditory tests provide the basis for defining dimension parameters of “noisiness”, i.e., parameters describing dimension-relevant features of a signal.

The SD carried out based on the NB-stimuli has shown that “noisiness” is highly correlated with two of the used antonym pairs, namely not noisy - noisy and not hissing - hissing [5].

The MDS of the NB-stimuli yielded ratings of all three dimensions for every stimulus. Within this test, in particular, two stimulus types obtained negative ratings of “noisiness” [5]:

- Stimuli distorted by background noise or circuit noise.
- Stimuli containing signal-correlated noise caused by the speech codec within the considered transmission system.

The auditory analysis of the N-stimuli has shown that the MOS of the N-stimuli, and thus, their ratings of the quality dimension “noisiness”, depend on the level and type of additive noise $n_{add}(k)$ (background noise, circuit noise, terminal noise) and the amount of signal-correlated noise, e.g., caused by the speech codec according to the ITU-T standard G.726 (Adaptive Differential Pulse Code Modulation, ADPCM) [7].

3.3. Dimension parameters of “noisiness”

The results of the auditory experiments led to the assumption that the following three signal parameters are relevant for the human perception and rating of the dimension “noisiness”:

- Intensity of additive noise: This parameter describes the quality-decreasing influence of additive noise that is contained in a signal $y(k)$ on the dimension “noisiness” due to the noise intensity, i.e., its loudness.
- Spectral composition of additive noise: This parameter indicates the quality-decreasing influence of additive noise on the dimension “noisiness” due to its spectral composition, i.e., its timbre.
- Amount of signal-correlated noise: This parameter characterizes the influence of the amount of signal-correlated noise on the quality dimension “noisiness” of a stimulus.

4. Dimension estimator

The dimension estimator for “noisiness” is based on a comparison of the speech signals $y(k)$ and $x(k)$, i.e., the signal that has been transmitted over the system under test and the respective original signal. The instrumental measure comprises three processing steps: (1) the preprocessing of the signals $x(k)$ and $y(k)$, (2) the extraction of the dimension parameters, and (3) the dimension model that estimates “noisiness” as a function of the respective dimension parameters.

4.1. Preprocessing

The joint preprocessing of the signals $x(k)$ and $y(k)$ serves two purposes: (1) the modelling of the listening situation in an auditory test, and (2) the compensation of differences between both signals that do not affect the human rating of “noisiness” in an auditory test. Thus, the preprocessing contains the limitation of the signal $x(k)$ to the frequency range of the transmission system under test and the filtering of both signals with a bandpass filter that models the frequency response of the handset used in the auditory test. Furthermore, a possible delay between the signals $x(k)$ and $y(k)$ is eliminated and their levels are adjusted to a common active speech level of -26 dBv. The preprocessed speech signals are denoted as $x'(k)$ and $y'(k)$.

4.2. Assessment of dimension parameters

During the assessment of the dimension parameters of “noisiness”, the signals $x'(k)$ and $y'(k)$ are processed in $L$ segments of length $M$ (32 ms). Each segment is weighted with a Hann window, $h_w(\kappa)$. Subsequent segments overlap by $\Delta M$ samples. Thus, for the segments $x'_w(\kappa,l)$ and $y'_w(\kappa,l)$, $\kappa = 0, 1, \ldots, M - 1$ and $l = 0, 1, \ldots, L - 1$, we have,

$$x'_w(\kappa,l) = x'(\kappa + l \cdot (M - \Delta M)) \cdot h_w(\kappa), \tag{1a}$$

$$y'_w(\kappa,l) = y'(\kappa + l \cdot (M - \Delta M)) \cdot h_w(\kappa). \tag{1b}$$

In the first step of the parameter assessment, a voice-activity detection is performed for the signal $x'(k)$ using a voice-activity detector (VAD) according to [8]. The output $vad_x(r)(l)$ of the VAD indicates whether the signal segment $x'_w(\kappa,l)$ contains speech activity or not. It is one for segments with speech activity and zero for segments without speech activity. The result of the VAD, $vad_x(r)(l)$, is used for the speech-pause classification of both signals. Thus, $vad_y(l) = vad_x(r)(l)$ holds.
4.2.1. Intensity of additive noise

In our work, the additive noise \( n_{add}(k) \) contained in the signal \( y'(k) \) is described by the signal \( y'(k) \) within speech pauses.

As a measure for the intensity of the estimated additive noise \( \hat{n}_{add}(k) \), the noise level \( N L_{\text{add}}(\mu) \) is used. It describes the mean energy of the estimated noise and is defined in [9]. The noise level \( N L_{\text{add}}(\mu) \) is calculated based on the power-density spectrum (PDS) of the signal \( \hat{n}_{add}(k) \). The PDS results from averaging the periodograms \( P_{y'\mu}(\mu, l) \) of segments \( y'_\mu(\kappa, l) \) without speech activity. Subsequent segments overlap by 50%.

Here, the parameter \( \mu \) indicates the discrete frequencies \( f_\mu \) where the PDS is evaluated:

\[
 f_\mu = f_S / M \cdot \mu,
\]

with \( f_S \) being the sampling frequency of the speech signals \( x'(k) \) and \( y'(k) \). In this work, \( f_S = 32 \) kHz is chosen. Thus, every speech segment \( x'_\mu(\kappa, l) \) and \( y'_\mu(\kappa, l) \) contains \( M = 1024 \) samples. The periodogram \( P_{y'\mu}(\mu, l) \) of a segment \( y'_\mu(\kappa, l) \) is defined as

\[
P_{y'\mu}(\mu, l) = \frac{1}{M} \sum_{n=0}^{M-1} y'_\mu(n, l) \cdot e^{-j 2\pi \mu n}.
\]

The PDS \( \tilde{n}_{\text{add}}(\mu) \) of the estimated noise \( \hat{n}_{\text{add}}(k) \) results to

\[
\tilde{n}_{\text{add}}(\mu) = \frac{1}{L_{y'}(\mu)} \sum_{(l|\text{vad}(l) = 0)} P_{y'\mu}(\mu, l).
\]

Here, the parameter \( L_{y'}(\mu) \) gives the number of frames \( y'_\mu(\kappa, l) \) without speech activity. The level \( N L_{\text{add}}(\mu) \) of the estimated noise is calculated as follows from the PDS \( \tilde{n}_{\text{add}}(\mu) \):

\[
N L_{\text{add}}(\mu) = |X'(\mu, l)| = \left| \text{DFT}\{x'_\mu(\kappa, l)\} \right|, \quad \text{for } l = 1 \text{ to } M - 1,
\]

\[
N L_{\text{add}}(\mu) = |Y'(\mu, l)| = \left| \text{DFT}\{y'_\mu(\kappa, l)\} \right|, \quad \text{for } l = 1 \text{ to } M - 1.
\]

Averaging the short-time magnitude spectra \( |X'(\mu, l)| \) and \( |Y'(\mu, l)| \) over such segments with speech activity yields the mean magnitude-spectra \( \bar{X}(\mu) \) and \( \bar{Y}(\mu) \):

\[
\bar{X}(\mu) = \frac{1}{L_{y'}(\mu)} \sum_{(l|\text{vad}(l) = 1)} |X'(\mu, l)|,
\]

\[
\bar{Y}(\mu) = \frac{1}{L_{y'}(\mu)} \sum_{(l|\text{vad}(l) = 1)} |Y'(\mu, l)|.
\]

The parameters \( L_{y'}(\mu) \) and \( L_{x'}(\mu) \) give the number of frames with speech activity in each signal. Here, segments without speech activity are disregarded as signal-correlated noise impairs the speech quality more strongly in segments with than in segments without speech activity.

The difference of the spectra \( \bar{X}(\mu) \) and \( \bar{Y}(\mu) \) is used for characterizing the total amount of noise, the signal transmission has effected in the signal \( y'(k) \):

\[
N_{\text{add}}(\mu) = \bar{Y}(\mu) - \bar{X}(\mu).
\]

The estimated spectrum \( N_{\text{add}}(\mu) \) includes both types of noise existent in signal segments with speech activity, signal-correlated noise and additive noise. Thus, an estimate for the signal-correlated noise present in the signal \( y'(k) \) is determined by subtracting an estimated spectrum \( N_{\text{add}}(\mu) \) of the additive noise from the spectrum \( N_{\text{add}}(\mu) \).

To allow for a comparison of the estimated additive noise with the estimated transmission noise, the spectrum \( N_{\text{add}}(\mu) \) is calculated similarly to the spectra \( \bar{X}(\mu) \) and \( \bar{Y}(\mu) \). It is determined by averaging the spectra \( |Y'(\mu, l)| \) over segments \( y'_\mu(\kappa, l) \) that are classified as speech pauses:

\[
N_{\text{add}}(\mu) = \frac{1}{L_{y'}(\mu)} \sum_{(l|\text{vad}(l) = 0)} |Y'(\mu, l)|.
\]
In this work, the spectrum of the estimated signal-correlated noise is defined as the difference of the spectra $N^{(tr)}(\mu)$ and $N^{(add)}(\mu)$ normalized to the mean spectrum $X(\mu)$:

$$N^{(cor)}_{spec}(\mu) = \frac{N^{(tr)}(\mu) - N^{(add)}(\mu)}{X(\mu)}. \quad (11)$$

The dimension parameter $N^{(cor)}$ is calculated as follows from the estimated spectrum $N^{(cor)}_{spec}(\mu)$:

$$N^{(cor)} = \max \left(0, \frac{1}{\Delta \mu} \sum_{\mu=\mu_1}^{\mu_3} N^{(cor)}_{spec}(\mu) \right). \quad (12)$$

In this process, only the spectral components between 3.4 kHz and 10 kHz to 15 kHz are used. A sufficient number of stimuli and choice of test conditions is required for an adequate number of stimuli and choice of test conditions.

In this process, only the spectral components between 3.4 kHz and 10 kHz to 15 kHz are used. A sufficient number of stimuli and choice of test conditions is required for an adequate number of stimuli and choice of test conditions.

In this process, only the spectral components between 3.4 kHz and 10 kHz to 15 kHz are used. A sufficient number of stimuli and choice of test conditions is required for an adequate number of stimuli and choice of test conditions.

4.3. Dimension model

Based on the quality ratings $\text{MOS}_N$ of the N-stimuli and these stimuli’s values of the dimension parameters of "noisiness", the following model for the quality estimates of the N-stimuli is determined by means of least-squares regression:

$$\text{MOS}_N = -1.165 - 0.072 \cdot \frac{NL^{(add)}}{\text{dBvp}} - 3.625 \cdot 10^{-4} \cdot f^{(add)} \quad (13)$$

$$-0.819 \cdot N^{(cor)} + 0.047 \cdot (N^{(cor)})^2.$$

As expected, all three parameters $NL^{(add)}$, $f^{(add)}$ and $N^{(cor)}$ have proven to be relevant for the instrumental assessment of the quality of the N-stimuli, and thus the dimension "noisiness". According to the quality model in Eq. (13) each dimension parameter possesses a vectorial influence on the estimates $\hat{\text{MOS}}_N$, i.e., the less the parameter values the better the quality. While the parameters $NL^{(add)}$ and $f^{(add)}$ both show a linear relation with the quality estimates, the parameter $N^{(cor)}$ shows a relation pursuant to a saturation curve with the estimates $\hat{\text{MOS}}_N$. This relation is modelled by means of a quadratic function; the values of $N^{(cor)}$ lie left of its vertex.

5. Reliability of dimension estimates

For training the quality model in Eq. (13) all 84 N-stimuli have been used. For these stimuli the quality estimates show a correlation of $\rho = 0.905$ with the ratings determined in the auditory test. The root mean-square error (RMSE) between the values $\text{MOS}_N$ and $\text{MOS}_N$ is $\text{RMSE} = 0.289$.

Within a crossvalidation it is analyzed how reliable the model in Eq. (13) estimates the quality of unknown test stimuli. Twenty different sets of training and test stimuli have been investigated. Each set of training stimuli $M_{Tr}$ contains 74 of the 84 N-stimuli. The respective stimulus set $M_{Ca}$ for checking the estimator includes the remaining ten N-stimuli. Each set $M_{Ca}$ comprehends five stimuli of each speaker and contains at least two stimuli with background noise, system noise (circuit and terminal noise) and signal-correlated noise. In this study, the estimates $\hat{\text{MOS}}_N$ show a mean correlation of $\bar{\rho} = 0.902$ (std. dev.: $\sigma_\rho = 0.020$) with the results of the auditory test. The mean RMSE between $\hat{\text{MOS}}_N$ and $\text{MOS}_N$ is given by $\text{RMSE} = 0.351$ (std. dev.: $\sigma_{\text{RMSE}} = 0.008$).

To sum up, a model according to Eq. (13) estimates the quality of unknown test stimuli with approximately the same reliability as for known training stimuli in case it is trained with an adequate number of stimuli and choice of test conditions.

6. Conclusion and outlook

Our work aims at the instrumental assessment and analysis of telephone-band speech quality based on measuring the underlying perceptual dimensions “directness/frequency content”, “noisiness”, and “continuity”. In the present paper, an instrumental estimator for the dimension “noisiness” was introduced. Three parameters were used, each describing a dimension-salient feature: the level $NL^{(add)}$ of additive noise and the center of gravity $f^{(add)}$ of its PDS as well as the intensity $N^{(cor)}$ of signal-correlated noise. These dimension parameters provide the basis for an instrumental measure for “noisiness”. Its results, the dimension estimates $\hat{\text{MOS}}_N$, show a correlation of $\rho > 0.95$ with the respective results of auditory tests.

Due to its performance, the instrumental measure for “noisiness” provides a good basis for a reliable dimension-based measure for the overall quality of telephone-band speech. While an estimator for “directness/frequency content” has been presented earlier work [6], the estimator for “continuity” is under development. In our future work, all three estimators will be combined to form a model for a system’s MOS. Furthermore, the reliability of the resulting quality model will be compared with the reliability of other speech-quality measures such as PESQ and TOSQA.

7. Acknowledgements

The authors would like to thank the “Deutsche Forschungsgemeinschaft” (DFG) for the financial support of the presented work under the grants HE 4465/1-1 and MO 1038/2-1.

8. References


