Toningenieur Projekt

Similarity Disturbance Quality - SiDiQ Improvements and Comparison to other Metrics

Felix Perfler
Matr. No.: 01637739

Electrical Engineering and Audio Engineering, Master's Program - UV 066 413
University of Music and Performing Arts Graz
Graz University of Technology

Supervisor: Univ.Prof. Dipl.-Ing. Dr.techn. Alois Sontacchi

Graz, June 22, 2022

Erklärung

Hiermit bestätige ich, dass mir der Leitfaden für schriftliche Arbeiten an der KUG bekannt ist und ich die darin enthaltenen Bestimmungen eingehalten habe. Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Felix Perfler, 01637739
Graz, den 22. Juni 2022

Unterschrift des Verfassers

Zusammenfassung

Die Entwicklung eines Maßes welches den Qualitätseindruck von Sprache wieder spiegelt ist nach wie vor ein aktives Forschungsfeld. Neu entwickelte Methoden basieren, im Gegensatz zu signalbasierten Ansätzen wie Perceptual Evaluation of Speech Quality (PESQ)[1], auf neuronalen Netzen, zum Beispiel DNSMOS [2]. Mit Similarity Disturbance Quality (SiDiQ) werden bei der Berechnung auch subjektive Bewertungen berücksichtigt.

Basierend auf bestehender Arbeiten von B. Stahl and A. Sontacchi [3] konnte die Genauigkeit der Metrik verbessert werden. Hierfür wurden die Bausteine des Modells beleuchtet und konnten teilweise vereinfacht werden. SiDiQ wurde mit einigen anderen populären Methoden zur Schätzung der Sprachqualität auf drei Datensätzen verglichen.

Die resultierende Metrik wurde in Form eines Python Modules implementiert und ist frei verfügbar[1]. Trotz der erzielten Verbesserungen stellt sich PESQ nach wie vor als verlässlichere Metrik heraus.

Abstract

The task to quantify the quality of speech signals is still an active field of research. In addition to signal based methods, such as Perceptual Evaluation of Speech Quality (PESQ)[1], models based neural network approaches, for example DNSMOS [2], are being developed. Similarity Disturbance Quality (SiDiQ) takes also subjective human ratings into account.

Based on the previous work by B. Stahl and A. Sontacchi [3] improvements to the performance could be made. Therefore the steps of the model were examined and some could be simplified. SiDiQ was compared to a number of other popular speech quality estimation algorithms on three different datasets.

The resulting metric was implemented using Python and the module is made freely available[1]. Even thought it was possible to improve the performance on the tested dataset, PESQ still proves to be the more reliable metric.

[1] `https://git.iem.at/stahl/sidiq`

Contents

Chapter 1

Introduction

Estimation of speech quality is still an active field of research. With the desire to build better speech enhancement algorithms, in addition to traditional applications like speech transmission, it is important to have a reliable metric, which models the properties of enhanced speech signals. Therefore, the properties of the human hearing system needs to be taken into account.

Over the years many such algorithms have been developed. One of the best known is Perceptual Evaluation of Speech Quality (PESQ) (see [1]), which takes one degraded and a reference signal to estimate the quality. With the growing popularity of neural networks, algorithms based on this technology are being developed. For example DNSMOS (see [2]), the name being derived from combining Deep Noise Suppression (DNS) with Mean Opinion Score (MOS), which estimates the quality without a reference. In contrast to computational quality assessment methods and the neural network approach, Similarity Disturbance Quality (SiDiQ) incorporates subjective ratings of test subjects in the computation.

First shown in [4], SiDiQ tries to predict the overall speech quality in a new novel approach, by echoing the human perception of speech. Given a stimulus and a clean speech reference signal the model aims to predict the quality of the stimulus. Using those signals two features are engineered. The first object tries to quantify the preservation of the target signal compared to the degraded one. The second object describes looks at the disturbance of background sounds present in the degraded signal. Next the computed features are used to estimate the overall quality, by utalising linear regression. The coefficients of the linear regressor were computed using results obtained by a istening experiment from [4]. During this listening experiment, participants were presented with a reference and degraded speech signal. They were then asked to rate the degraded signal in terms of overall quality, “preservation of the target signal”, and “disturbance by background sounds” [4]. All the parameters of the resulting model are tuned using cross-validation grid search.

The model proposed in [3] was changed by removing the saliency model from the disturbance feature computation and replacing the k-nearest-neighbour regression by a linear regression resulting in a better performance.

This work is based on previous publications, see [3] and [4]. The performance of the proposed model could be improved. By examining the existing model a number of simplifications are introduced.In addition to removing the saliency model from the disturbance feature computation, the k-nearest-neighbour regression was replaced by a linear regression.

Finally, the metric is also compared to different existent speech quality assessment metrics. PEMO-Q [5], using the implementation from [6], fwsegSNR [7], SI-SDR [8], Kastner's 2f-model, in its MATLAB implementation [9], PESQ, using the implementation of [7] modified following [10], and ViSQOl in the speech flavor ([11, 12]).

Additionally, a new larger dataset, with higher variability in terms of available signals, was used to determine how well the metric generalises. It can be shown that on this dataset SiDiQ is only outperformed by PESQ and DNSMOS, performing equally well as ViSQOL-speech.

Chapter 2

Similarity Disturbance Quality - SiDiQ

The model can be summarized by the following graphic:

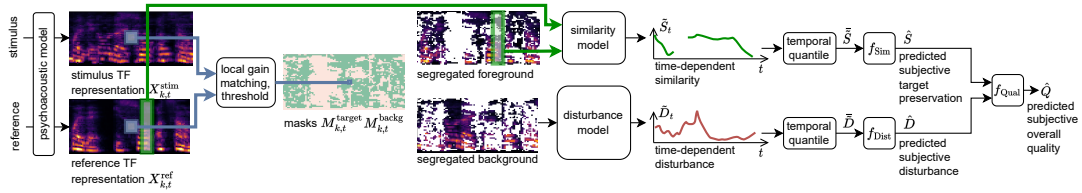


Figure 2.1: Proposed model, graphic adapted from [3]

In short a stimulus and reference signal are treated with a psychoacoustic model. After calculation of the Short Time Fourier Representation (STFT) a foreground and background mask is estimated. Using the resulting segregated foreground and background objects features are engineered. Using subjective ratings and applying linear regression those features are in return mapped to the overall quality using the same method. For the linear regression result from a listening experiment were used.

2.1 Loudness Model

In order to take the human perception of loudness into account, two different loudness models, the one used by PESQ and Zwicker's model, were implemented and tested. In the end, the loudness model as implemented in PESQ, showed better results and was therefore made the default when computing the metric.

2.1.1 PESQ Loudness Model

In order to obtain the PESQ-equivalent spectra the pre-processing steps of the PESQ model are applied to the signals. These steps perform perceptual transformations and are implemented according to [13]. The reference and stimulus signal are aligned to a standard listening level through filtering in order to model a telephone handset. Once the signals are level aligned the perceived loudness can be calculated. First the spectrum $X_{k,t}$ is computed using Short Time Fourier Transform (STFT) using a Hamming window w and 32ms frame lenght with a 50% overlap.

$$X_{k,t} = \sum_{n=-\inf}^{\inf} x[n]w[n-t]e^{-jkn} \tag{2.1}$$

Next the power spectrum transformed to a modified Bark one using a filterbank with 42 band channels via a multiplication in the Fourier domain. To obtain the loudness spectrogram $S_{k,t}$, the bark spectrogram $B_{t,k}$ is transformed to:

$$S_{k,t} = s_l \cdot \left(\frac{P_0(k)}{0.5}\right)^{\gamma} \cdot \left[\left(0.5 + 0.5\frac{B_{t,k}}{P_0(k)}\right)^{\gamma} - 1\right] \tag{2.2}$$

with s_l being a loudness scaling factor and $P_0(k)$ the absolute hearing threshold for the k-th Bark band. The factor γ is set to 0.23. Lastly, all frequency bins below a frequency dependent threshold are set to zero.

2.1.2 Zwicker Loudness Model

The loudness spectra according to the Zwicker model [14] can be obtained as follows. First the signals are normalized to a sound pressure level of 85 dB full scale range (FSR). Next the spectrogram is computed using a Hann window, frame length of 32ms and 50% overlap. The computed power spectrum is than transformed to a Bark one with 42 bins. The loudness spectrogram $S_{k,t}$ is then computed as followed given the Bark spectrogram $B_{t,k}$:

$$S_{k,t} = 0.08 \cdot \left(\frac{E_k}{E_0}\right)^{0.23} \cdot \left[\left(0.5 + 0.5 \cdot \frac{B_{t,k}}{E_k}\right)^{0.23} - 1\right] \tag{2.3}$$

with E_0 being the squared reference pressure level p_0 as $2 \cdot 10^{-5}$ and E_k the excitation level at the absolute hearing threshold.

2.2 Feature Extraction

Using the time-frequency representations of the reference $X_{k,t}^{ref}$ and the stimulus $X_{k,t}^{stim}$ a mask is calculated, which segregates the background from the foreground. In order to decide whether the frequency bin should be counted as foreground or background, an equalization gain $g_{k,t}^{ref}$ is calculated according to:

$$g_{k,t}^{ref} = \frac{\sum_{\tilde{k}=-\frac{\alpha}{2}}^{\frac{\alpha}{2}} \sum_{\tilde{t}=-\beta+\gamma}^{\gamma} X_{k+\tilde{k},t+\tilde{t}}^{stim} X_{k+\tilde{k},t+\tilde{t}}^{ref}}{\sum_{\tilde{k},\tilde{t}} X_{k+\tilde{k},t+\tilde{t}}^{ref^2}} \tag{2.4}$$

The patch used for the computation is defined by the frequency patch width α, patch length β and look ahead parameter given as γ. Applying the computed gain to the reference representation, a ratio can be computed as:

$$R_{k,t} = \frac{X_{k,t}^{stim}}{g_{k,t}^{ref} \cdot X_{k,t}^{ref}} \tag{2.5}$$

The foreground mask is than calculated by passing the bin values to a modified sigmoid function. The center of the sigmoid function represents a threshold value, whether to count the bin as foreground or background. The background mask is computed as follows:

$$M_{k,t}^{backg} = 1 - M_{k,t}^{foreground} \tag{2.6}$$

2.2.1 Disturbance Feature

The time dependent disturbance feature is calculated as the ratio of the instantaneous background loudness and the total mean overall loudness.

$$\tilde{D}_t = \frac{L_t^{\text{backg}}}{\bar{L}^{\text{total}}} \tag{2.7}$$

The background loudness over time is calculated as:

$$L_t^{\text{backg}} = \frac{1}{\tau_{\text{loud}}} \sum_{k=1}^{K} \sum_{\tilde{t}=-\tau_{\text{loud}}}^{0} v_k M_{k,t+\tilde{t}}^{\text{backg}} X_{k,t+\tilde{t}}^{\text{stim}} \tag{2.8}$$

And the the overall loudness is computed thus:

$$\bar{L}^{\text{total}} = \sum_{\tilde{k},\tilde{t}} X_{k+\tilde{k},t+\tilde{t}}^{\text{stim}} \tag{2.9}$$

2.2.2 Similarity Feature

The similarity feature time series can be calculated by first computing the weighted correlation coefficient of the segregated target and the reference.

$$\rho_t^{\text{target}} = \frac{\sum_{k=1}^{K} \sum_{\tilde{t}=-\beta}^{0} v_k M_{k,t+\tilde{t}} X_{k,t+\tilde{t}}^{\text{stim}} X_{k,t+\tilde{t}}^{\text{ref}}}{\sqrt{\sum_{k,\tilde{t}} v_k (M_{k,t+\tilde{t}} X_{k,t+\tilde{t}}^{\text{stim}})^2 \sum_{k,\tilde{t}} v_k X_{k,t+\tilde{t}}^{\text{ref}^2}}} \tag{2.10}$$

This coefficient corresponds to the similarity feature as follows:

$$\tilde{S}_t = \begin{cases} \rho_t^{\text{target}}, & \text{if } \bar{X}_t^{\text{ref}} > \zeta \\ \text{undefined}, & \text{else} \end{cases}, \tag{2.11}$$

where $\bar{X}_t^{\text{ref}}$ is the mean reference loudness calculated as:

$$\bar{X}_t^{\text{ref}} = \frac{1}{\beta} \sum_{k=1}^{K} \sum_{\tilde{t}=-\beta}^{0} v_k X_{k,t+\tilde{t}}^{\text{ref}} \tag{2.12}$$

and ζ is a parameter chosen to be the minimum reference loudness for which a reliable similarity can be considered (see table 2.1).

2.2.3 Overall Quality

Once both the time-dependent similarity and disturbance is computed, weighted quantile values of those time series are extracted (see figure 2.2).

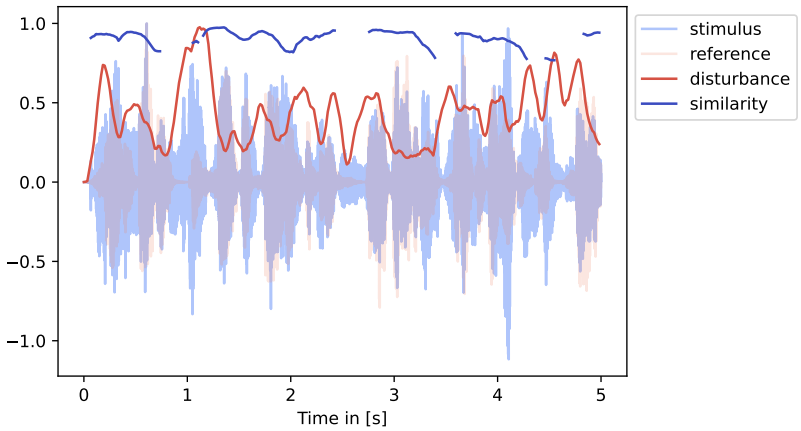


Figure 2.2: Exemplary calculated features time series

First the signal is windowed, using a window centered at the quantile value. The weighted quantile is than calculated according to 2.13, where w is the chosen window of length N,

t is the ordered time series and q is the quantile.

$$q_{weighted} = \frac{\sum_{n=q-N/2}^{q+N/2} t[n] \cdot w[n]}{\sum_{n=0}^{N} w[n]} \tag{2.13}$$

The window is of type Hann with a length of 10 samples. Using linear regression those quantiles of the features $\tilde{S}$ and $\tilde{D}$ values are mapped onto $\hat{S}$ and $\hat{D}$ of the subjective quality aspects. The results are once again mapped onto the overall rated subjective quality using linear regression.

2.3 Model Parameter Tuning

The model parameters are tuned using grid search using cross validation to find the best combination. First, for each possible parameter combination, the feature time series $\tilde{S}$ and $\tilde{D}$ are computed. Then every possible quantile value for each time series are calculated. For the cross-validation, one participant and mixture is excluded. Next the prediction of the mean ratings on the stimuli of the excluded mixture are computed using linear regression fitting on the mean ratings of all the other stimuli in the remaining mixtures of all remaining participants. In order to obtain the optimal parameters the mean squared errors (MSEs) for all the the cross validation folds is calculated. Next the mean MSEs of each excluded mixture is averaged, therefore computing the global MSE. The parameter set with the lowest global MSE is then the optimum (see Table 2.1).

Table 2.1: Model parameters

Disturbance model parameters	
parameter	values
masking threshold	{1.1, **1.3**}
masking threshold width	{0.001, **0.2**}
max reference gain	{1.1, 1.3, **1.5**}
patch frequency width α	{**5**, 7, 9} bands
patch length β	{**0.192**} s
lookahead γ	{**0.032**} s
surprise time constant	{**0.6**}
surprise hopsize	{**0.016**}
surprise exponent	{**0.4**}
minimum information threshold	{**1.1**}
feature quantile	{0.0, 0.1, 0.2, **0.4**, 0.5}
second feature quantile	{0.6, **0.7**, 0.8, 0.9, 1.0}
Similarity model parameters	
parameter	values
minimum information threshold ζ	{10, 20, 30, 40, 50, **60**}
feature quantile	{**0.0**, 0.1, 0.2, 0.4, 0.5}
second feature quantile	{**0.6**, 0.7, 0.8, 0.9, 1.0}

Chapter 3

Results

The algorithm as described in chapter 2 was implemented in Python using the Tensorflow framework [15]. Since it performed superior the PESQ equivalent loudness model (as described in section 2.1.1) was used. The model parameters were set to the one found in the grid search (see table 2.1). For each reference and stimulus pair the SiDiQ metric was computed. Additionally, if available the two individual features, similarity and disturbance, were calculated as well.

SiDiQ was evaluated on a number of different dataset to measure its performance against other popular speech quality assessment metrics. In addition to the training dataset the same dataset as in [3], a modified version of the SEBASS dataset [16] was used, and the dataset from [17] was used. In order to measure the performance, correlation coefficients between the subjective ratings and the model predictions were computed. The correlation coefficient matrix $R_{i,j}$ can be computed, given the covariance matrix C as:

$$R_{i,j} = \frac{C_{i,j}}{\sqrt{C_{i,i} \cdot C_{j,j}}} \tag{3.1}$$

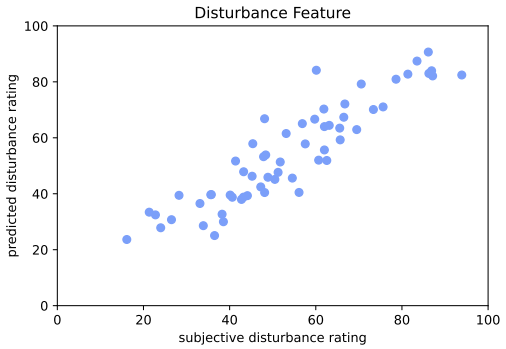
3.1 Training Dataset

The training dataset consists of twenty stimuli drawn from the PAESS dataset, as found in [18]. Contained in the dataset are five target/interferer(s) scenarios and their clean speech target signals. The stimuli are treated with four different speech enhancement algorithms. Derived from eleven more mixtures, additionally forty-four more stimuli were created, containing signals degraded by office noise and stimuli from the ChiME-4 challenge [19]. Both traditional and deep learning based source seperation algorithms were used to process the data. Ratings were obtained trough a webMUSHRA [20] listening experiment. Twenty-six participants rated the stimuli in terms of “preservation of the target signal”, “disturbance by background sounds”, and "overall quality". The dataset is comprised of

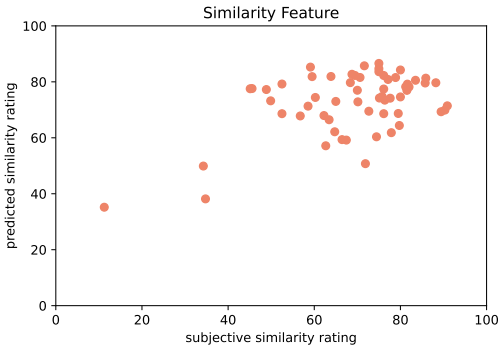
60 rated stimuli. During model fitting artificial references were introduced, yet for the following tables, those ratings are ignored.

Table 3.1: Correlation coefficients of SiDiQ applied for the training dataset.

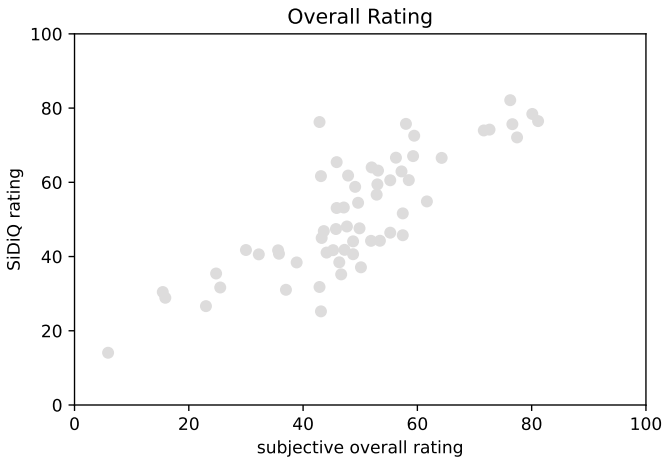
aspect	correlation coefficient
background rating	0.91
similarity rating	0.50
overall rating	0.82



(a) Background rating subjective compared to predicted



(b) Similarity rating subjective compared to predicted



(c) Overall rating subjective compared to predicted

Figure 3.1: Features and overall ratings

Table 3.2: Correlation coefficients of different metrics.

metric	correlation coefficient
SiDiQ	0.82
SI-SDR	0.81
fwsegSNR	0.72
PESQ	**0.87**
PEMO-Q	0.79
Kastner	**0.87**
Visqol Speech	0.77
DNSMOS	0.51

3.2 SEBASS Dataset

As was done in the previous work SiDiQ was evaluated on a modified SEBASS dataset. Since the task requires speech signals all music was removed from the dataset. Moreover, the PEASS sub-dataset, as well as ratings on anchor and hidden reference signals were not considered either. The PEASS sub-dataset was removed as it was used for model fitting. Therefore, the dataset consists of 224 speech stimuli. Those stimuli are drawn from eight mixture scenarios processed by 28 different source seperation algorithms. All audio examples were downsampled to 16 kHz from the original rate of 48 kHz. The dataset does not contain similarity and disturbance ratings, therefore only the overall rating can be compared for different metrics.

Table 3.3: Correlation coefficients of different metrics.

metric	correlation coefficient
SiDiQ	**0.80**
SI-SDR	0.34
fwsegSNR	0.48
PESQ	0.73
PEMO-Q	0.66
Kastner	0.67
Visqol Speech	0.65
DNSMOS	0.59

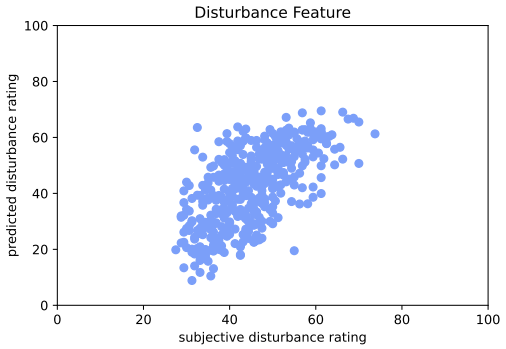
3.3 NOIZEUS Dataset

The dataset contains 30 IEEE sentences, spoken by three male and female speakers respectively, resulting in 5 sentences per speaker. The speech signals were then corrupted using eight different real-world noises with different signal to noise ratios. The resulting signals were all downsampled to a sampling frequency of 8kHz. Additionally, references and stimuli were filtered to simulate a telephony handset. This is done by applying a modified Intermediate Reference System (IRS) same as ITU-T P.862 [21]. The resulting signals were treated with 13 different speech enhancement algorithms.

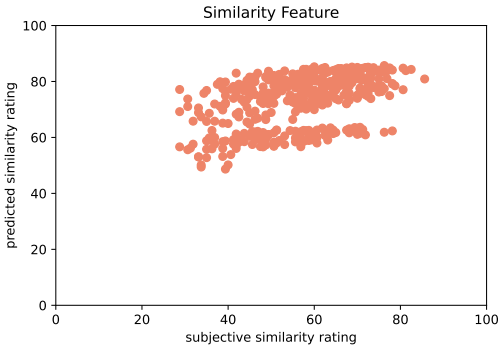
Below, the correlation coefficients for the similarity, disturbance, and overall rating are listed (see table 3.4.) Moreover, table 3.5 compares the ratings of different metrics.

Table 3.4: Correlation coefficients of SiDiQ applied to the NOIZEUS dataset.

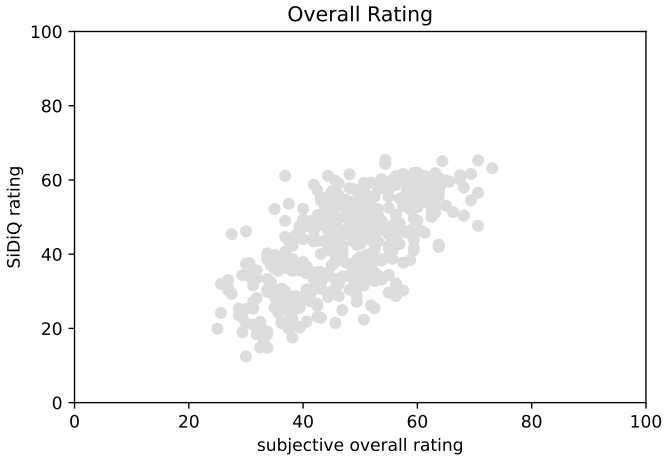
aspect	correlation coefficient
disturbance rating	0.65
similarity rating	0.45
overall	0.67



(a) Background rating subjective compared to predicted



(b) Similarity rating subjective compared to predicted



(c) Overall rating subjective compared to predicted

Figure 3.2: Features and overall ratings

Table 3.5: Correlation coefficients of different metrics.

metric	correlation coefficient
SiDiQ	0.67
SI-SDR	0.38
fwsegSNR	0.57
PESQ	**0.81**
PEMO-Q	0.53
Kastner	0.59
Visqol Speech	0.67
DNSMOS	0.71

Chapter 4

Conclusion and Outlook

Even though the computation was simplified the model performance on the more varied NOIZEUS dataset was improved. The previous model, described in [4], achieved a correlation coefficient of 0.4, when comparing the subjective and objective rating of the overall speech quality. It can be shown that the simplified model achieves a higher score of 0.67 (see table 3.4.)

Nonetheless, the performance of SiDiQ can still be improved. One of the simplifications was to remove the saliency computation, when calculating the disturbance feature. Introducing an alternative to the saliency could prove fruitful, since the aim of the saliency, to model the capability of the background model to draw a listener's attention, should still be a useful objective in order to draw conclusion on speech quality. Looking at table 3.1 shows that the fitting of the disturbance feature to the training dataset performes well, as the correlation coefficient is high. The coefficient for the similarity on the other hand is lower. Therefore, conducting a new listening experiment could potentially result in a training dataset which captures this aspect better, hence force resulting in a potential performance gain of the similarity feature.

With the growing popularity of neural networks for speech enhancement tasks, the usage of SiDiQ as a cost function could be considered, as adapting speech quality metrics to be used in that context can work [13].

The implementation of SiDiQ, as well as the optimal parameters found in the grid search are freely available[1].

[1] `https://git.iem.at/stahl/sidiq`

Bibliography

[1] A.W. Rix et al. “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs”. In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*. Vol. 2. 2001, 749–752 vol.2. DOI: 10.1109/ICASSP.2001.941023.

[2] Chandan K A Reddy, Vishak Gopal, and Ross Cutler. *DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors*. 2020. DOI: 10.48550/ARXIV.2010.15258. URL: https://arxiv.org/abs/2010.15258.

[3] Benjamin Stahl and Alois Sontacchi. “SIDIQ: Computational Quality Assessment of Enhanced Speech Based on Auditory Figure-Ground Segregation, Similarity, and Disturbance”. In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2021, pp. 96–100.

[4] Benjamin Stahl and Alois Sontacchi. “Speech enhancement quality assessment based on aspect-specific qualities: A preliminary analysis”. In: *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE. 2021, pp. 351–355.

[5] Rainer Huber and Birger Kollmeier. “PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception”. In: *IEEE Transactions on audio, speech, and language processing* 14.6 (2006), pp. 1902–1911.

[6] Emmanuel Vincent. “Improved perceptual metrics for the evaluation of audio source separation”. In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer. 2012, pp. 430–437.

[7] Philipos C Loizou. *Speech enhancement: theory and practice*. CRC press, 2007.

[8] Jonathan Le Roux et al. “SDR–half-baked or well done?” In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2019, pp. 626–630.

[9] Peter Kabal et al. “An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality”. In: *TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University* (2002), pp. 1–89.

[10] ITU-T Recommendation. “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”. In: *Rec. ITU-T P. 862* (2001).

[11] Andrew Hines et al. “ViSQOL: an objective speech quality model”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), pp. 1–18.

[12] Michael Chinen et al. “ViSQOL v3: An open source production ready objective speech and audio metric”. In: *2020 twelfth international conference on quality of multimedia experience (QoMEX)*. IEEE. 2020, pp. 1–6.

[13] Juan Manuel Martin-Donas et al. “A deep learning loss function based on the perceptual evaluation of the speech quality”. In: *IEEE Signal processing letters* 25.11 (2018), pp. 1680–1684.

[14] Eberhard Zwicker and Bertram Scharf. “A model of loudness summation.” In: *Psychological review* 72.1 (1965), p. 3.

[15] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[16] Thorsten Kastner and Jürgen Herre. “An efficient model for estimating subjective quality of separated audio source signals”. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2019, pp. 95–99.

[17] Yi Hu and Philipos C Loizou. “Subjective comparison of speech enhancement algorithms”. In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 1. IEEE. 2006, pp. I–I.

[18] Valentin Emiya et al. “Subjective and objective quality assessment of audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2046–2057.

[19] Emmanuel Vincent et al. “An analysis of environment, microphone and data simulation mismatches in robust speech recognition”. In: *Computer Speech & Language* 46 (2017), pp. 535–557.

[20] Michael Schoeffler et al. “webMUSHRA—A comprehensive framework for web-based listening tests”. In: *Journal of Open Research Software* 6.1 (2018).

[21] ITU-T. *Recommendation 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, 2001.