Audio quality predictions based on power and envelope power SNRs

Thomas Biberger, Stephan D. Ewert

Medizinische Physik and Cluster of Excellence Hearing4All, University of Oldenburg, Email: thomas.biberger@uni-oldenburg.de

Introduction

There are various reasons, e.g. speech enhancement or saving storage/transmission capacity, why music or speech signals are processed by algorithms. Often such signal processing introduces distortions to the original signal, which might be hardly perceivable or might affect the audio quality if they become detectable. For development and evaluation of such algorithms, instrumental perceptual quality models are beneficial as they can partly replace time consuming and expensive listening tests. Here, an approach for a reference-based audio quality model applicable to a wide range of signal distortions is proposed based on the recent speech intelligibility model by Jørgensen et al. [1]. The proposed model calculates power and envelope-power signal-to-noise ratios (SNR_{dc}, SNR_{env}) between the manipulated and the original signal as features on multiple time scales. Those features are calculated for the case of an increment (where the algorithm adds energy to the original signal) as well as the case of a decrement (where the algorithm removes energy), based on this ability of humans to detect changes in both directions. Both measures are combined and mapped to a continuous quality rating scale. Audio quality predictions are compared to perceptual quality ratings using three different databases from the literature.

Model description

In the following, the proposed multi-resolution, multi-channel envelope power spectrum model is termed mr-mcEPSM. The mr-mcEPSM is based on [1] and combines properties of the envelope power spectrum model (EPSM, [2]) as well as the power spectrum model (PSM, [3]). The model inputs are processed and unprocessed signals, whereby processed signals are referred to as test signals while unprocessed signals are referred to as reference signals.

Front-End

The first stage of the mr-mcEPSM contains an outer- and middle-ear filter followed by a 4th-order Gammatone filterbank with one-ERB [4] bandwidth and third-octave spacing from 63 to 15000 Hz, representing the auditory filters. In each auditory channel, the envelope of the filtered signal is extracted via Hilbert transformation and filtered by a 1st-order low-pass with a cut-off frequency of 150 Hz ([2]). Next, for SNR_{env} calculation, the envelope of each auditory channel is filtered by a modulation filterbank with bandpass filters ranging from 2 to 256 Hz, which are parallel to a 3rd-order low-pass filter with a cut-off frequency of 1 Hz. The subsequent multi-resolution stage divides the output of the modulation filters into temporal segments with a duration corresponding to the inverse of the center frequency of the specific modulation filter. Thus, low modulation filters supply a low temporal resolution, whereas higher modulation filters supply a high temporal resolution. Subsequently, the ac-coupled envelope power for all temporal segments is calculated and the $SNR_{env,m,n,i}$ between the processed and the unprocessed signal is derived, whereas m, n, and i refer to a specific auditory channel, modulation channel, and temporal segment, respectively. Averaging $SNR_{env,m,n,i}$ across temporal segments results in $SNR_{env,m,n}$, which represents the 2-dimensional front-end output with the dimensions modulation and auditory (center) frequency.
For $SNR_{dc,m}$ calculation, the intensity within each of the auditory channels m is calculated for the processed and unprocessed signal.
The front-end output provides modulation information by $SNR_{env,m,n}$ as well as intensity information by $SNR_{dc,m}$ which is in the following referred to as SNR. The SNR represents a 2-dimensional matrix composed of m auditory and n+1 intensity/modulation channels.

Back-End

In contrast to psychoacoustic and speech intelligibility predictions [5], where only an increase of intensity or modulation power in the processed signal compared to the reference (increment) is considered, audio quality predictions additionally consider the decrement case. Increments and decrements in the processed signal compared to the unprocessed signal can generally occur and thus can be perceived (e.g., [6]) and can affect the quality judgements. The 2-dimensional increment and decrement SNR-matrices are combined by taking the maximum of each of the entries. The SNRs of the resulting matrix are combined across auditory and intensity/modulation channels to obtain a single SNR value. The SNR value has then to be mapped to a continuous quality rating scale applying a logarithmic function with a lower and upper limit to the model output. Hereby, the lower limit reflects the minimum of audio quality degradation which can be resolved by subjects, while the upper limit reflects the deviation for which maximum audio quality degradation is perceived.

Evaluation

As presented in [5] at last year's DAGA conference, the mr-mcEPSM accounts for various psychoacoustic experiments as just-noticable differences (JND) in intensity, non-simultaneous masking, simultaneous masking, hearing threshold, amplitude-modulation (AM) detection, AM-discrimination, and AM-masking. Hereby, the

performance of the mr-mcEPSM can be compared to the established perception model (PEMO, [7]) showing that the signal features derived by the front-end cover the basic aspects of perception. Furthermore, it was shown in [5] that the mr-mcEPSM in combination with a speech intelligibility back-end can account for speech intelligibility in a wide variety of background noises (stationary, fluctuating and reverberation) similar to the original model by Jørgensen et al. [1].

In order to examine the mr-mcEPSM's predictive power for audio quality assessment, three databases with different kinds of distortions were used in this study. An objective quality measure was derived by applying the mr-mcEPSM including the back-end for audio quality to the unprocessed and processed signals. The Pearson correlation coefficient r between subjective ratings and model predictions (objective quality measure) was calculated and used as performance measure. This measure is an indicator for the linear dependence between objective and subjective quality ratings, where r=1 means a total linear dependence, while r=0 indicates no linear dependence. The mr-mcEPSM results are compared with performance measures from the three objective (model-based) approaches PEMO-Q [8], CASP-Q [9], and HASQI [10], which were taken from a study by Harlander et al. [9].

Audio Codecs

The Audio Codec database (for details see [8]) results from six subjective evaluations for several low-bit-rate audio codecs, carried out by the International Telecommunication Union (ITU) and the Moving Pictures Experts Group (MPEG). The database consists of 433 test items, based on speech as well as music material. Test signals are mainly non-linear distorted. Subjective quality ratings from 19 to 91 expert listeners were obtained by using the Subjective Difference Grade (SDG), which ranges from -4 (very annoying) to 0 (imperceptible). The mr-mcEPSM predictions showed a correlation coefficient r=0.8, while PEMO-Q, CASP-Q and HASQI yielded r values of 0.9, 0.82 and 0.6.

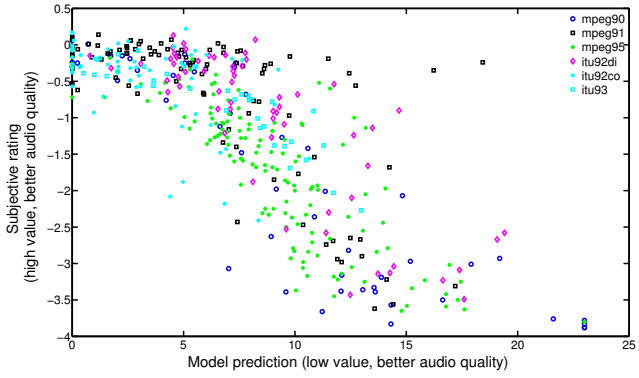

Figure 1: Prediction results of mr-mcEPSM for the Audio Codec database. The abscissa shows model predictions, while the ordinate indicates subjective ratings.

Noise Reduction

This database, established by Hu and Loizou [11] consists of 1920 speech items, resulting from 16 speech samples mixed with four different types of background noise at two SNR levels, processed by 15 different noise reduction algorithms (Spectral Subtraction, Wiener Filter, etc.). This kind of signal processing leads also to non-linear distortion. The subjective evaluation was carried out by 40 normal hearing (NH) subjects via the Mean Opinion Score (MOS), which ranges from 1 (bad) to 5 (excellent). For this database, mr-mcEPSM, PEMO-Q and HASQI reached similar performance, indicated by r values of 0.85, 0.85 and 0.86, while CASP-Q showed the best performance reflected by r=0.91.

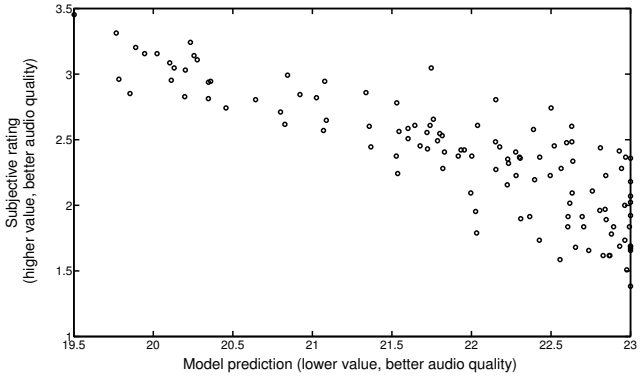

Figure 2: Audio quality predictions of mr-mcEPSM for the Noise Reduction database. The abscissa shows model predictions, while the ordinate indicates subjective ratings.

Audio Source Separation

The Audio Source Separation database from Emiya et al. [12] has 70 test items. These items consist of speech as well as music material. In order to generate the test signal, the reference signal containing only a target signal (e.g. target speaker) is superimposed by an interfering signal (e.g. interfering talker). The aim of the algorithm is to suppress the interfering signal and to restore the original target signal. This processing also results mainly in non-linear distortions. Subjective results are based on the MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA, [13]) paradigm, which was carried out by 23 NH subjects. Here, mr-mcEPSM achieved a correlation coefficient r=0.8 and thus outperforms PEMO-Q, CASP-Q and HASQI with r values of 0.63, 0.68 and 0.54.

Summary and conclusion

An auditory model (mr-mcEPSM) was suggested, consisting of a psychoacoustically motivated front-end and a task dependent back-end for either predicting psychoacoustic experiments and speech intelligibility [5] or audio quality. In terms of audio quality, it has been shown that the model provides stable prediction results (r>=0.8) across different kinds of non-linear distortions. In contrast to other highly specialized quality models extraordinary good performance for a specific class of distortions is

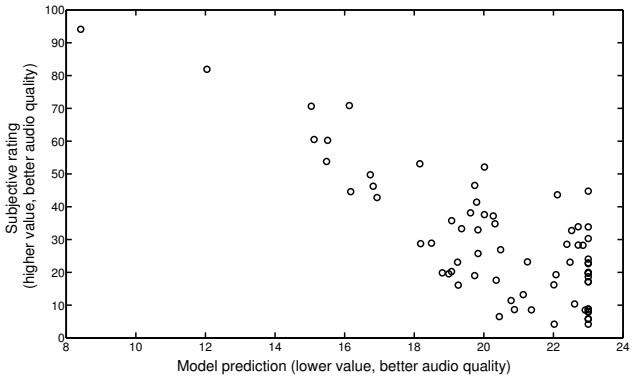

Figure 3: Results of quality predictions for the Audio Source Separation database by mr-mcEPSM. The abscissa shows model predictions, while the ordinate indicates subjective ratings.

not achieved without further (database dependent) tuning of the model. In future work, the monaural modeling approach presented in this study should be combined with a binaural model (e.g., [14]) to account for monaural as well as binaural features.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, FOR 1732, TP E).

References

[1] Jørgensen, S., Ewert, S. D and Dau, T.:. A multi-resolution envelope-power based model for speech intelligibility. J. Acoust. Soc. Am. **134** (2013), 436-446

[2] Ewert, S. D. and Dau, T:. Characterizing frequency selectivity for envelope fluctuations. J. Acoust. Soc. Am. **108** (2000), 1181-1196

[3] Patterson, R. D. and Moore, B. C. J. :. Auditory filters and excitation patterns as representations of frequency resolution, chap. Frequency selectivity in hearing, 123-177. Academic Press

[4] Moore, B. C. J. and Glasberg, B. R.:. Suggested formulae for calculating auditory filter bandwidth and excitation patterns. J. Acoust. Soc. Am. **74** (1983), 750-753

[5] Biberger, T., Ewert, S.D.: Psychoacoustic, speech intelligibility, and audio quality predictions based on envelope power SNRs. Fortschritte der Akustik - DAGA 2014, Oldenburg, Deutschland, Dega e.V., Berlin.

[6] Oxenham, A.J.:. Increment and decrement detection in sinusoids as a measure of temporal resolution. J. Acoust. Soc. Am. **102** (1997), 1779-1790

[7] Dau, T, Püschel, D. and Kohlrausch, A.: A quantitative model of the "effective" signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. **99** (1996), 3615-3622

[8] Huber, R. and Kollmeier, B.:. Pemo-q - a new method for objective audio quality assessment using a model of auditory perception. IEEE **14** (2006), 1902-1911

[9] Harlander, N., Huber, R., Ewert, S.D.: Sound Quality Assessment Using Auditory Models. J. Audio Eng. Soc. **62** (2014), 324-336

[10] Kates, J.M., Arehart, K.H.: The Hearing-Aid Speech Quality Index (HASQI). J. Audio Eng. Soc. **58** (2010), 324-336

[11] Hu, Y., Loizou, P.C.: A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. **122** (2007), 1777-1786

[12] Emiya, V., Vincent, N., Harlander, N., Hohmann, V.: Subjective and Objective Quality Assessment of Audio Source Separation. IEEE Transaction on Audio, Speech, and Language Processing **19** (2011), 2046-2057

[13] ITU: ITU-R Recommendation BS.1534-1: Method for the assessment of intermediate quality levels of coding systems. 2003

[14] Fleßner, J., Ewert, S.D., Kollmeier, B.,and Huber, R.: Spatial audio quality prediction using a binaural auditory model.. International Hearing Aid Research Conference (IHCON), Lake Tahoe, 13.-17. August 2014.