Psychoacoustic, speech intelligibility, and audio quality predictions based on envelope power SNRs

Thomas Biberger, Stephan D. Ewert
Medical Physics, University of Oldenburg, Germany, Email: thomas.biberger@uni-oldenburg.de

Introduction

There are many factors as room properties, environmental noise or the transmission path corrupting clean speech or audio signals in realistic listening situations. As measures for the impact of these factors, speech intelligibility, speech or audio quality are widely used. Detectability of changes in a stimulus can be assessed by psychoacoustic experiments. In order to predict quality and speech intelligibility measures, as well as psychoacoustic data within a single auditory model framework, a measure dependent decision stage (back-end) and a common auditory front-end, providing perceptually relevant features of speech or audio signals, are required. Here, an approach is proposed based on the recent speech intelligibility model by Jørgensen et al. [1] which extended an earlier amplitude-modulation processing model [2]. The proposed model front-end calculates power- and envelope-power signal-to-noise ratios ($\mathrm{SNR}_{\mathrm{dc}}$, $\mathrm{SNR}_{\mathrm{env}}$) as features on multiple time scales.

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 the unprocessed signal can be regarded as the reference signal. In case of psychoacoustic experiments, the processed signal contains the target stimulus and the reference does not.

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 100 to 12000 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 $\mathrm{SNR}_{\mathrm{env}}$ 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 $\mathrm{SNR}_{\mathrm{env,m,n,i}}$ 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 $\mathrm{SNR}_{\mathrm{env,m,n,i}}$ across temporal segments results in $\mathrm{SNR}_{\mathrm{env,m,n}}$, which represents the 2-dimensional front-end output with the dimensions modulation and auditory (center) frequency.

For $\mathrm{SNR}_{\mathrm{dc,m}}$ 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 $\mathrm{SNR}_{\mathrm{env,m,n}}$ as well as intensity information by $\mathrm{SNR}_{\mathrm{dc,m}}$ and is in the following denoted as SNR. The SNR represents a 2-dimensional matrix composed of m auditory and n+1 intensity/modulation channels.

Back-End

For predicting psychoacoustic experiments, the SNRs are combined across auditory and intensity/modulation channels to obtain a final single value SNR. The final SNR is compared to a threshold criterion, that requires a SNR-value higher than -6 dB for detecting the processed signal, based on [5].

For predicting speech intelligibility a back-end as proposed in [6] was used. Hereby, the SNR is transformed into a percentage of correct responses of presented speech material.

In contrast to psychoacoustic and speech intelligibility predictions, 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. This is done, because both increments and decrements in the processed signal compared to the unprocessed signal can be perceived 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. Then 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 by applying a non-linear fitting as shown in [7] to the model output.

Evaluation

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. The performance of the mr-mcEPSM can be compared to the established perception model (PEMO, [8]) showing that the signal features derived by the front-end cover the basic aspects of perception.

Speech Intelligibility

To test whether the mr-mcEPSM as a modified version of the original model mr-sEPSM [1] conserves its abilities in predicting speech intelligibility, Danish speech material from the Conversational Language Understanding Evaluation (CLUE) in combination with stationary (speech-shaped noise; SSN), fluctuating (AMSSN, International Speech Test Signal), and reverberation interferer as it is used in [1], were tested. The model input signals were noise interferer alone (unprocessed) and the noise interferer plus speech (processed) at a certain SNR, which finally results in a predicted percentage of correct responses as a function of the SNR. Based on this function, the point of 50% correct responses, termed speech reception threshold (SRT), was derived for each interferer condition. Here, the root mean squared error (RMSE) between SRTs derived from experimental results and model prediction across all speech intelligibility experiments was used as a performance measure. The RMSE-value for the original mr-sEPSM is 1.9 dB, while the RMSE-value for the suggested model mr-mcEPSM corresponds to 1.4 dB.

Audio quality

In order to examine the mr-mcEPSM's predictive power for audio quality ratings, 433 audio files from mpeg- and itu-databases (see [7]), which were processed by different low bit-rate audio codecs, are utilized. An objective quality measure was derived by applying the mr-mcEPSM including the back-end for audio quality to the unprocessed and processed (by audio codecs) signals. Subjective ratings are given by the subjective difference grade (SDG), which reflects quality ratings between the unprocessed and processed signals. The SDG ranges from -4 (very annoying impairment) to 0 (imperceptible impairment). The pearson correlation coefficient r is used to reflect the matching between objective predictions with subjective ratings and has a value of r = 0.8 for the mr-mcEPSM. This can be considered as a mediocre prediction performance. In comparison to the mr-mcEPSM, the established PEMO-Q ([7]) achieves a significantly higher prediction performance of r = 0.9.

Summary and conclusion

An auditory model (mr-mcEPSM) was suggested, consisting of a single front-end and a measure dependent back-end. The model accounts for several psychoacoustic effects while maintaining the predictive power of the underlying approach [1] for speech intelligibility. The model appears generally suited to predict audio quality. However the more complex PEMO-Q shows a higher prediction performance. In future studies, further audio databases should be evaluated by the mr-mcEPSM to get more insight in the capabilities and limitations of this model.

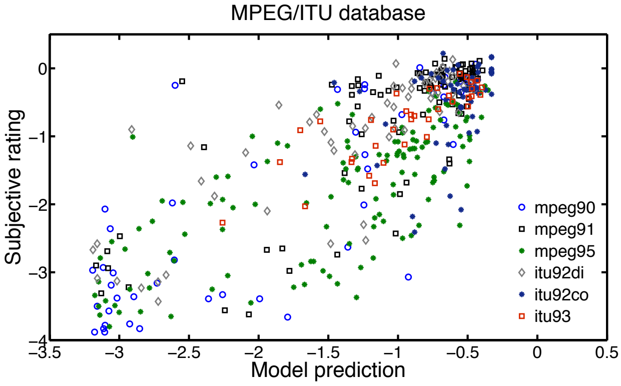

Figure 1: Results of audio quality predictions by mr-mcEPSM. Abscissa shows model predictions, while the ordinate indicates subjective ratings.

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] Ewert, S. D. and Dau, T:. External and internal limitations in amplitude-modulation processing. J. Acoust. Soc. Am. **116** (2004), 478-490

[6] Jørgensen, S. and Dau, T:. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. **130** (2011), 1475-1487

[7] 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

[8] 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