QUALITY ASSESSMENT OF MULTI-CHANNEL AUDIO PROCESSING SCHEMES BASED ON A BINAURAL AUDITORY MODEL

Jan-Hendrik Fleßner[†] *Stephan D. Ewert*[⋆] *Birger Kollmeier*[⋆] *Rainer Huber*[†]

[†] HörTech gGmbH, Oldenburg, Germany, and Cluster of Excellence "Hearing4All"
[⋆] Medizinische Physik, Universität Oldenburg, and Cluster of Excellence "Hearing4All"

ABSTRACT

A perceptual, binaural audio-quality model is introduced. The model was developed for predicting any kinds of perceived spatial quality differences between two audio signals in multi-channel reproduction and audio processing schemes. It employs a recent binaural auditory model as front-end to provide perceptually relevant binaural features for the reference and test audio signal. Correlations between the binaural features of both signals are combined to an overall spatial quality measure by the use of multivariate adaptive regression splines (MARS). Furthermore, a database was generated to train and evaluate the model. The database contains various multi-channel audio signals, which were subjectively assessed in formal listening tests with 15 trained listeners. The results show different model prediction performances depending on the type of quality degradation. Combination of the proposed spatial quality measure with established monaural quality measures improved the predictive power.

Index Terms— Spatial audio quality, binaural auditory model, objective quality assessment

1. INTRODUCTION

To evaluate the audio quality of audio processing or reproduction schemes, formal listening tests still represent the "gold standard". However, disadvantages of these subjective tests are the high investments of time and money. Consequently, much effort has been put into the development of computational models for audio quality prediction as a tool to complement listening tests, e.g. during the development phase of audio processing algorithms. Most of the up-to-date models are oriented towards the human perception, e.g. [1, 2]. These models typically compute "internal representations" in higher stages of the auditory system for a reference and a test audio signal. Similarity or distance measures are then used to compare these internal representations and to compute a value reflecting the mean subjective opinion on the perceived quality of the test signal relative to the reference signal.

To date, multi-channel audio systems have been well established and are widely used. Different techniques for recording, processing, transmission/coding and reproduction of spatial audio can affect the perceived spatial quality of the reproduced audio in various ways. On the other hand, rather few audio quality models include spatial quality prediction. The audio quality model recommended by ITU-R recommendation BS.1387 ("PEAQ") [3] is able to process stereo audio signals, but each channel is handled separately. Efforts were made to extend PEAQ with a spatial processing stage, e.g. [4, 5]. The binaural auditory perception model of Breebaart *et al.* [6], which is a binaural extension of the perception model "PEMO" of Dau *et al.* [7], is used as front-end in [8] to predict aspects of room acoustics perception based on interaural time differences. In addition to the psychoacoustic motivated models, Rumsey *et al.* [9] developed a model for predicting spatial transmission and reproduction audio quality. For this purpose, spatial attributes were identified and defined. These attributes were assigned to suitable technical measures.

In the strict sense, all comparison-based models do not predict quality directly, but the perceived overall difference or similarity between a given pair of audio signals. It is up to the user of the model to interpret the model output appropriately, e.g., in terms of (relative) quality.

The aim of the current study was to develop a perceptual model for predicting the magnitude of the perceived overall spatial difference. The recent binaural auditory model of Dietz *et al.* [10] was chosen as front-end. To ensure a general applicability of the proposed quality model, specialization for a single task during development and training had to be avoided. Thus, signals which contain various types of spatial quality alterations were generated. Subjective quality assessment was performed for the test signal database, and the results were used for training and evaluation of the proposed quality model.

One fundamental problem is the separation of purely spatial (binaural) signal distortions and monaural distortions as consequence of spatial signal manipulation and as consequence of potential difficulty of listeners to clearly separate them. Therefore, the binaural quality measures extracted in the proposed approach were combined with established monaural measures [1], [2].

The generation of the database, the model structure and the evaluation of the binaural quality model will be presented in the following.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. SUBJECTIVE SPATIAL AUDIO QUALITY DATA

Given that the binaural quality model is to predict spatial quality differences only, monaurally perceptible quality differences (e.g. timbre) should ideally not be contained in the signal database or should at least be small compared to spatial quality differences. This would also help in the listing tests, because listeners would not have to differentiate between spatial and non-spatial quality aspects. Accordingly, a database consisting of subjectively assessed multi-channel audio signals was generated.

2.1. Signal generation

Multi-channel recordings of speech (two samples), music (three samples) and nature sound (one sample) were convolved with five stereo room impulse responses and added up to down-mix the signals to stereo signals. The stereo room impulse responses were measured in a circular 5-loudspeaker setup (radius 1.2 m) with three loudspeakers in front and two in the back of a pair of microphones (cardioid characteristic) in ORTF [11] stereo configuration. In the reference condition, the loudspeakers were positioned at 0°, ±45° and ±135°. Six different methods were used to introduce spatial quality alterations of these reference signals. It was ensured that the alterations differ considerably from each other, while choosing algorithms of practical relevance. Three of them manipulated the stereo room impulse responses by (1) using partly different responses, i.e. from different loudspeaker positions, (2) changing the inter-microphone time or level differences, and (3) changing the direct-sound-to-reverberation ratio. The other three methods were experimental hearing aid algorithms: (4) an interaural phase jitter to increase the apparent source width and to decrease localization acuity, (5) binaural noise reduction attenuating signal parts with low interaural coherence [12], (6) a hearing aid algorithm [13] for noise reduction, for which speech samples from the Oldenburg Sentence Test [14] mixed with different noise signals were used as input signals. This algorithm was a binaural beamformer that employs speech-distortion-weighted multichannel filtering. Here, the amount of noise reduction was reduced by adjusting the RMS value of the noise in the processed signal to match that of the unprocessed signal. The overall effect on the perceived quality was supposed to be dominated by an altered spatial impression. Overall level differences between each test signal and corresponding reference signals were compensated. The database covers 136 test signals altogether.

2.2. Subjective signal assessment

The subjective assessment was realized by pairwise comparisons of test and reference signals. In contrast to other typical subjective audio quality assessment methods, where subjects are instructed to interpret any perceivable difference between a test and reference signal as quality degradations of the test signal, here the task for the listeners was to quantify the perceived overall difference between the two presented signals on a numerical rating scale ranging from 0 (“no difference”) to 4 (“obvious difference”) in 0.1-steps. Overall difference ratings were interpreted as quality differences afterwards. This made rating for the listeners more straightforward and easier. Apart from the two attributes at the ends of the rating scale, no intermediate attributes were used. The test procedure ran fully automatically in Matlab on a standard Windows PC, listeners entered their responses using a graphical user interface. Signals were played in a loop. Listeners were free to listen as long as they liked and they could switch back and forth quasi-instantaneously between reference and test signal at any time. Additionally, signal segments (> 300 ms) could be freely selected and played in a loop if wanted.

The measurements took place in a sound-attenuated booth. Audio signals were sampled at 44.1 kHz. (Reference and test signals processed by the hearing aid algorithms, however, had an original sample rate of 16 kHz). They were played back via external soundcard (RME Fireface UC) and headphones (Sennheiser HD650). 15 normal hearing listeners with former experience in listening tests participated in the study. Before the beginning of the actual tests, they performed a training session.

3. MODEL STRUCTURE

The spatial audio quality model consists of a front-end that extracts binaural features and a back-end that analyzes and correlates the binaural features and finally combines them to an overall spatial quality measure.

3.1. Extraction of binaural features

The front-end (Fig. 1) is based on the binaural auditory model of Dietz *et al.* [10]. A 500-Hz to 2-kHz second-order band-pass filter is used to approximate the outer/middle ear frequency response measured by Puria *et al.* [15]. The frequency-dependent behavior of the basilar membrane is simulated by a linear, fourth-order gammatone filterbank [16, 17] with center frequencies from 236Hz to 12978Hz. One gammatone filter is used per ERB (equivalent rectangular bandwidth, [18]) with, in contrast to [10], a bandwidth of 1.5 ERB, to account for increased effective binaural auditory filter bandwidths (e.g., [19]). The basilar membrane compression is modeled by instantaneous compression with an exponent of 0.4. A subsequent halfwave-rectification is applied to simulate the mechano-electrical transduction process in the inner hair-cells. After these steps, depicted with the block *peripheral processing* in Fig. 1, the binaural feature extraction follows: Two complex-valued gammatone filters [17] are applied to assess temporal fine-structure (fine-structure

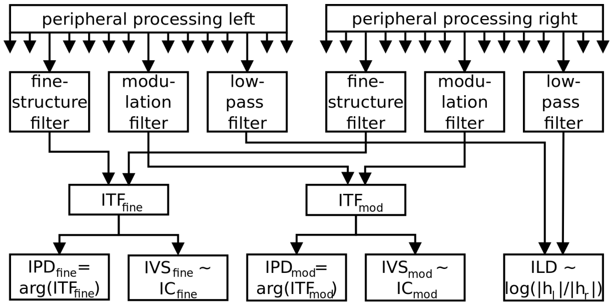

Fig. 1. Schematic diagram of the binaural feature extraction of the front-end.

filter) and envelope (modulation filter). The fine-structure filter applies the same center frequency as the respective basilar membrane filter and has a q-value of three. The modulation filter (first-order) extracts amplitude modulations with a center frequency of 100 Hz for each channel. For basilar membrane filters with center frequencies $\leq$ 1.14 kHz, the fine-structure filter is used, otherwise the modulation filter.

The interaural transfer function (ITF) is calculated from the complex signals of the fine-structure and modulation filter:

$$ITF(t) = y_l(t) \cdot y_r^*(t) = A_l(t) \cdot A_r(t) \cdot e^{j(\phi_l(t) - \phi_r(t))} \quad (1)$$

(with $A_l(t)$ and $A_r(t)$ being the instantaneous amplitude and $\phi_l(t)$ and $\phi_r(t)$ the instantaneous phase)

The ITF is low-pass filtered:

$$ITF_{LP}(t) = \int_0^\infty ITF(t-\tau) \cdot e^{-\tau/\tau_s} d\tau \quad (2)$$

The time constant τ_s is five times the cycle duration of the respective auditory filter, so the resolution of the 1 kHz-band is 5 ms. The interaural level difference (ILD) is computed by applying a low-pass filter in parallel to the fine-structure and modulation filter and computing the energy ratio of the left and right low-pass filter output. From the ITF, the interaural phase difference (IPD) and interaural vector strength (IVS) [10] can be derived. The IPD is the argument of ITF_{LP}. As described in [10], the IPD can be ambiguous and the ILD can be used to dissolve this.

To obtain the interaural time difference (ITD), the IPD is devided by the mean instantaneous frequency of the left and right signal. The IVS is computed from the ITF, providing a measure similar to interaural coherence:

$$IVS(t) = \frac{\left|\int_0^\infty ITF(t-\tau) \cdot e^{-\tau/\tau_s} d\tau\right|}{\int_0^\infty \left|ITF(t-\tau) \cdot e^{-\tau/\tau_s}\right| d\tau} \quad (3)$$

3.2. Transformation of binaural features into a quality measure

The back-end processes IVS(t, f) and ITD(t, f) of the reference and the test signal (Fig. 2). Both binaural features are segmented into 300-ms time frames and further processed frame by frame with 50 % overlap.

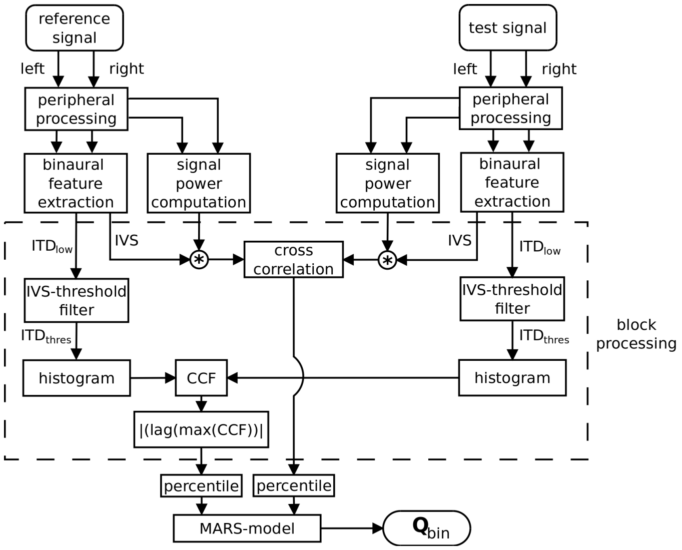

Fig. 2. Schematic diagram of the binaural spatial quality model.

Assuming that quality differences are perceived more intensively if the audio signal power in the respective time-frequency section is higher, the IVS is weighted accordingly with the signal power computed in each basilar membrane filter using a moving 10-ms window. The first intermediate measure (monitoring the perceptual diffusiveness of the sound) is the linear cross correlation coefficient between the weighted IVS matrix of the reference and the test signal. The IVS is further used as a binary weighting for the ITD$_{low}$ (ITD of the 11 lowest frequency channels), selecting ITD values for time-frequency points with an IVS of 0.96 or higher (ITD$_{thres}$), (see [10, 20]). With these ITD "glimpses" likely originating from a distinct source, a histogram with the boundaries -1 ms and 1 ms and a bin width of 50 μs is built, providing an image of the spatial distribution of distinct sources. The cross correlation function (CCF) of the ITD histogram of the reference and the test signal is computed; the absolute value of the lag of the CCF maximum represents the second intermediate measure.

The 43th percentile of the IVS correlation sequence and the 86th percentile of the ITD measure sequence are combined to an overall spatial quality measure Q_{bin} by the use of multivariate adaptive regression splines (MARS) [21].

3.3. Combination of binaural and monaural quality measures

In order to also cover non-spatial, e.g. timbre differences, the above Q_{bin} measure was complemented by two established monaural quality measures: the linear distortion mea-

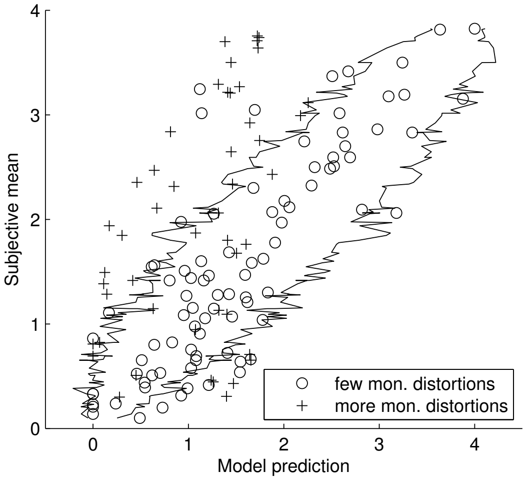

Fig. 3. Predictions of the binaural spatial quality model. Crosses: Signals with a higher amount of monaural cues. Black lines: Subjective 99%-confidence interval.

sure [1] and the PSM_t measure [2]. Another MARS model was trained to derive a combination of these three measures to an overall quality predictor.

4. EVALUATION

As a first test to assess if the proposed approach is in principle capable of predicting audio signal differences, the database described earlier was used to compare subjective ratings with corresponding model predictions. Due to the rather small size of the database, it was not separatated into "training" and evaluation sets. For the purely binaural approach, the MARS model was trained with the 88 signals of the database containing no or rather few residual monaural differences between test and reference, such as changes in timbre. The advantage of the MARS model is that it is highly restrictable and controllable. Consequently, the trained model could be checked for plausibility. Besides, the risk of over-fitting is rather low with this method. Figure 3 shows a scatter plot of mean subjective ratings from the listening test against the corresponding binaural quality model predictions. The 88 signals containing mostly spatial quality differences are indicated as circles, whereas crosses represent the remaining 48 signals of the database containing a higher amount of monaural variations. The linear cross correlation between subjective ratings and model predictions is 0.83 for the 88 signals and 0.66 for all signals. Moreover, 65 of the 88 quality predictions are located within the 99 %-confidence intervals of the subjective assessments, but only 14 of the 48 predictions. Figure 4 shows the results obtained with the combined binaural and monaural quality model. The MARS model was trained with the whole database. The linear cross correlation coefficient for all predictions increased to 0.79. The number of outliers of the 99 %-confidence region decreased from 57 to 44.

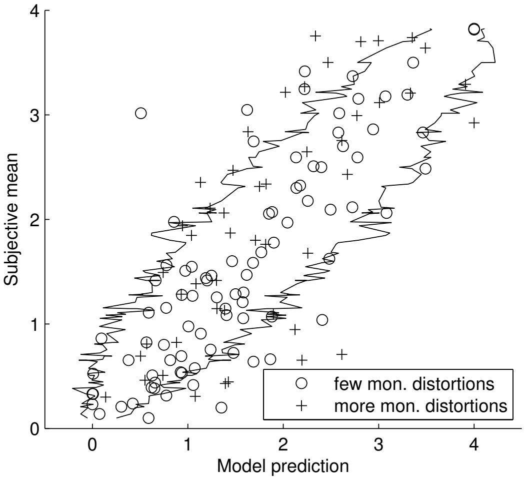

Fig. 4. As Fig. 3, but with results obtained with the overall quality predictor including binaural and monaural quality measures.

5. DISCUSSION AND CONCLUSION

The results obtained for the proposed binaural quality model are promising and suggest the validity of the general approach, i.e. to employ a perceptual, binaural auditory model to compute and compare binaural features of a pair of binaural audio signals. Very different types of spatial quality differences could be predicted by the model reasonably well, although similar high correlations with subjective ratings as often observed in monaural audio and speech quality assessment could not yet be achieved. Additional, larger databases will be required for extensive training/optimization and evaluation with separated training and validation data sets. It also became clear that purely spatial, i.e. binaurally perceivable differences between audio signals represent rather special cases of audio quality. Accordingly, the proposed combined binaural/monaural approach including both quality dimensions could improve the predictive power of objective quality assessment for spatial, multi-channel audio. In conclusion, the current study suggests that any model of overall audio quality for multi-channel audio should include (intermediate) measures for both types of quality differences.

6. ACKNOWLEDGEMENTS

This work was supported by the "Deutsche Forschungsgemeinschaft" DFG FOR1732, TPE.

7. REFERENCES

[1] B.C.J. Moore and C.-T. Tan, “Development and validation of a method for predicting the perceived naturalness of sounds subjected to spectral distortion,” *J. Audio Eng. Soc*, vol. 52, no. 9, pp. 900–914, 2004.

[2] R. Huber and B. Kollmeier, “PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[3] ITU-R, “Method for objective measurement of perceived audio quality,” 1998a, Series BS: Broadcast Services Recommendation BS.1387, International Telecommunications Union.

[4] Jeong-Hun Seo, Sang Bae Chon, Keong-Mo Sung, and Inyong Choi, “Perceptual objective quality evaluation method for high quality multichannel audio codecs,” *J. Audio Eng. Soc*, vol. 61, no. 7/8, pp. 535–545, 2013.

[5] M. Schäfer, M. Bahram, and P. Vary, “An extension of the PEAQ measure by a binaural hearing model,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8164–8168.

[6] J. Breebaart, S. van de Par, and A. Kohlrausch, “Binaural processing model based on contralateral inhibition. i. model structure,” *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1074–1088, 2001.

[7] T. Dau, D. Püschel, and A. Kohlrausch, “A quantitative model of the ’effectiv’ signal processing in the auditory system: I. model structure,” *J. Acoust. Soc. Am.*, vol. 99, pp. 3615–3622, 1996.

[8] J. van Dorp Schuitman, D. de Vries, and A. Lindau, “Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model,” *J. Acoust. Soc. Am.*, vol. 133, no. 3, pp. 15721585, 2013.

[9] F. Rumsey, S. Zielinski, P.J.B. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Meares, “QESTRAL (part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener,” in *Proc. 125th AES Conv., San Francisco CA*, Oct. 2008.

[10] M Dietz, S.D. Ewert, and V. Hohmann, “Auditory model based direction estimation of concurrent speakers from binaural signals,” *Speech Communication*, vol. 53, pp. 592–605, 2011.

[11] Thomas Görne, *Mikrofone in Theorie und Praxis*, vol. 8, Elektor-Verlag, 2007.

[12] G. Grimm, V. Hohmann, and B. Kollmeier, “Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1408–1419, 2009.

[13] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, “Theoretical analysis of binaural multimicrophone noise reduction techniques,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[14] K. Wagener, T. Brand, and B. Kollmeier, “Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests (Development and evaluation of a German sentence test part I: Design of the Oldenburg sentence test),” *Z. Audiol.*, vol. 38, pp. 4–15, 1999.

[15] S. Puria, W.T. Peake, and J.J. Rosowski, “Sound-pressure measurements in the cochlear vestibule of human-cdaver ears,” *J. Acoust. Soc. Am.*, vol. 101, pp. 2754–2770, 1997.

[16] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammetone function,” 1987, Paper presented at a meeting of the IOC Speech Group on auditory modeling at RSRE, 14-15 December.

[17] V. Hohmann, “Frequency analysis and synthesis using a gammatone filterbank,” *acta acustica united with acustica*, vol. 88, pp. 433–442, 2002.

[18] B.C.J. Moore and B.R. Glasberg, “Suggested formulae for calculating auditory filter bandwidths and excitation patterns,” *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[19] I. Holube, M. Kinkel, and B. Kollmeier, “Binaural and monaural auditory filter bandwidths and time constants in probe tone detection experiments,” *J. Acoust. Soc. Am.*, vol. 104, no. 4, pp. 2412–2425, 1998.

[20] C. Faller and J. Merimaa, “Source localization in complex listening situations: selection of binaural cues based on interaural coherence,” *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[21] J.H. Friedman, “Multivariate adaptive regression splines,” *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, 1991.