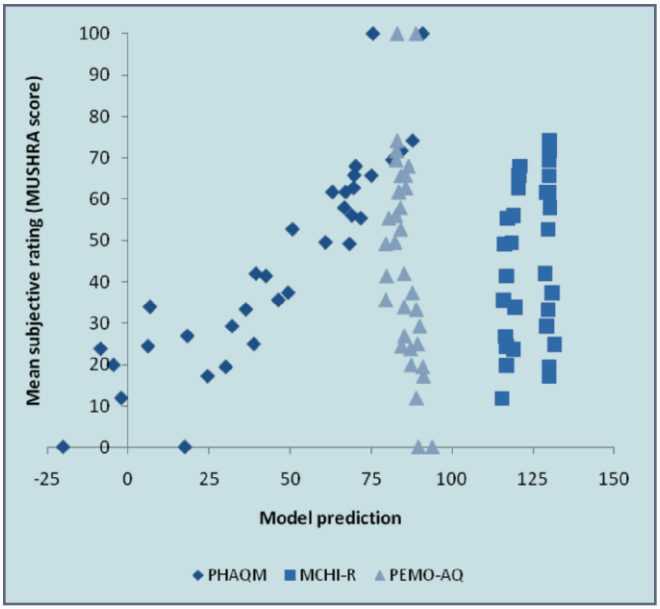

Fig. 2. Relation between overall quality predicted by three models and corresponding subjective ratings for frequency manipulation (normal-hearing listeners only)

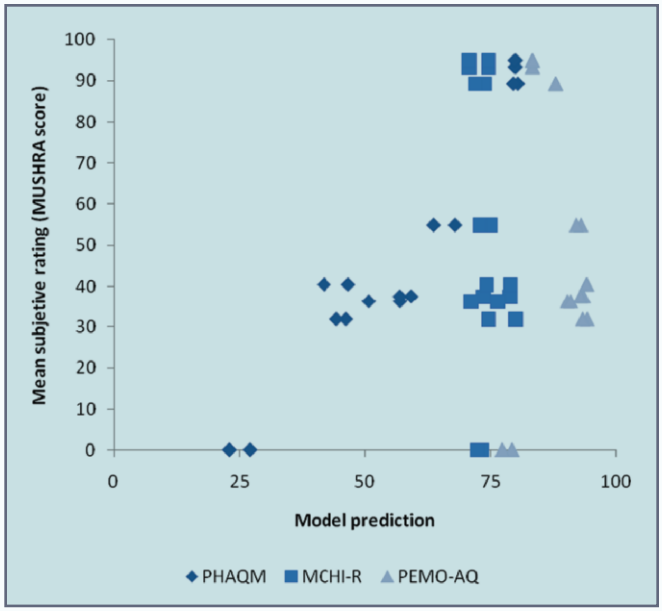

Fig. 3. Quality-prediction results for three models on the noise-reduction stimuli (hearing-impaired subjects)

models could be generalized to encompass alternative types of audio processing, and for this reason the data represented the effects of three different forms of frequency manipulation, a noise-reduction process, and various severe forms of filtering, clipping, and compression derived from Bramslow's OSSQAR data set. (OSSQAR was the name given to a predictive quality model he developed during an earlier Ph.D. project. The model predicted a range of quality attributes including loudness and sharpness.) The frequency manipulation involved variants of a nonlinear frequency compression process in which important high-frequency audio information was shifted downwards in the spectrum, to a more audible region for those with high-frequency hearing loss.

The models tested were PHAQM (based on the Perceptual Audio Quality Measure, developed by Beerends and Stemerdink), PEMO-AQ (based on a perceptual similarity model known as PSMt, developed by Huber), and MCHI-R (based on MCHI, a model to predict "pleasantness" for hearing-impaired persons, developed by Schmalfuss and Haubold). They had all been developed for use with hearing aids and impairment as part of a project organized by Hörtech, a German research institute specializing in the field. The general structure of the models is shown in Fig. 1. A test signal is fed into the test system (hearing aid or audio process) and the outputs of a normal-hearing and a hearing-impaired auditory model are compared. Objective measures are derived from the comparison output, which are used in the prediction of subjective ratings. Each model had been trained on MUSHRA-scaled audio quality data from listening tests undertaken by hearing-impaired listeners who listened to six commercial hearing aids of different price and manufacturer. Nine different stimuli included speech with and without background noise, bird song, music, and noisy environments. In the validation test described here, the stimuli used to gather the data sets employed had been a range of music and speech samples, with and without background noise.

As shown in Fig. 2, which depicts the results for one of the frequency manipulations using normal-hearing listener data, the PHAQM model gave predicted results that quite closely matched the subjective ratings of the stimuli, while the other two models did not perform well. The MCHI-R model performed moderately well on music but poorly on speech. Fig. 3 shows the results for the noise-reduction data set using hearing-impaired listener ratings. The correlation between predicted and actual ratings was again good for PHAQM, but the authors point out that this was improved by the clear extremes defined by the low- and high-anchor stimuli. In the middle of the range the results were more ambiguous and the subjective data had quite large intersubject differences. The other two models again appeared unable to predict the quality ratings given by subjects. With the OSSQAR data set, for the normal-hearing listeners, the PHAQM model performed best on the overall quality dimension for both signal types (music and speech). MCHI-R was best at predicting loudness with music stimuli, and PEMO-AQ best with speech. PHAQM performed best at predicting sharpness. For the hearing-impaired listeners, PHAQM performed best for overall quality and loudness on the music stimuli, whereas PEMO-AQ performed best on speech. None of the models produced particularly accurate predictions on speech in this test, however, and although music worked better the results were still a long way from being reliable.

The authors suggest that the PHAQM model came out as a clear winner for the data sets tested, although it did not perform well on the OSSQAR speech stimuli. Particularly interesting was that it seemed to work equally well for both hearing-impaired and normal-hearing listeners. They attribute some of the inadequacies demonstrated in these experiments to the fact that the models were calibrated to predict "big effects" such as the effect of different hearing