Users' Guide:
HASPI Version 2
HASQI Version 2
HAAQI Version 1

James M. Kates

Department of Speech Language and Hearing Sciences
University of Colorado
409 UCB, Boulder, CO 80309
James.Kates@colorado.edu

3 September 2013
Revision 1: 10 December 2015
Revision 2: 7 February 2020

Abstract

This report describes MATLAB functions that implement a new speech intelligibility index, a new speech sound quality index, and a new music quality index. The intelligibility index, HASPI version 2, replaces the previous version 1. The speech quality index, HASQI version 2, replaces the previous version of HASQI. The music quality index is HAAQI version 1. The new indices are based on a series of perceptual experiments involving normal-hearing and hearing-impaired listeners, and the indices represent minimum root mean-squared error fits of mathematical models to the subject data. The new intelligibility and quality indices share an auditory model that serves as a "front end" for the calculations. The auditory model incorporates the middle ear, auditory filters, cochlear dynamics, neural firing rate adaptation, and the auditory threshold found in the normal or impaired ear. The intelligibility index measures the degree of fidelity in reproducing the speech envelope combined with a measure of signal correlation that reflects the accuracy with which the signal temporal fine structure is reproduced. The quality indices also combine a measure of envelope fidelity with a measure of temporal fine structure fidelity, and also incorporate measures of changes in the long-term signal spectrum. Instructions are provided for using the MATLAB function for intelligibility prediction and the function for quality prediction. Predictions from the new indices are compared with predictions from the previous versions for noisy speech, and the limitations of the new indices are discussed.

1. Introduction

This memorandum describes how to use the HASPI version 2 intelligibility index and the HASQI version 2 and HAAQI version 1 quality indices. The HASPI version 2 speech intelligibility index (Kates and Arehart, 2020) replaces the previous HASPI version 1 (Kates and Arehart, 2014a) and coherence speech intelligibility indices (CSII) (Kates and Arehart, 2005). The HASQI version 2 speech quality index (Kates and Arehart, 2014b) replaces the previous version of HASQI (Kates and Arehart, 2010). The HAAQI music quality index (Kates and Arehart, 2015b) is a new index fit to previously-acquired data. HASPI version 2 offers accuracy comparable to that obtained for version for speech degraded by additive noise, peak clipping, and center clipping distortion, and for speech processed through a noise vocoder. It gives improved accuracy for speech signals modified by frequency lowering, noise suppression, and reverberation. HASQI version 2 gives accuracy comparable to that for the previous version for speech degraded by additive noise, peak clipping and center clipping, multichannel dynamic-range compression, and noise suppression. It gives improved accuracy for speech signals modified by frequency lowering, imperfect acoustic feedback cancellation, and noise vocoder reproduction. The HAAQI music index is accurate for music degraded by additive noise and babble, noise suppression, dynamic-range compression, and linear spectral changes.

2. Auditory Model

The new intelligibility and quality indices share a common auditory “front end” that is used to model the auditory periphery for normal and impaired hearing (Kates, 2013). Both indices compare the output of the auditory model for a degraded signal with the output for an unprocessed input signal. The model presented in this paper is an extension of the Kates and Arehart (2010) auditory model.

A detailed description of the new auditory model is presented in Kates (2013) and is summarized here. The overall index block diagram is presented in Fig 1. The comparison of the processed and reference signals requires that they be temporally aligned, so part of the model is the temporal alignment of the signals. The first alignment step is a broadband signal alignment. Each signal then goes through the middle ear and cochlear mechanics models, after which the delay of the processed signal in each frequency band is adjusted to maximize the cross-correlation with the reference signal in that band. The separate signals then go through the inner hair-cell (IHC) model, followed by compensation for the group delays of the auditory filters. In the final processing step the auditory model outputs are converted into the signal features used to compare the processed signal with the reference signal. The basilar membrane (BM) vibration signal in each frequency band is compressed using the same control function as for the envelope in that band, so the envelope of the BM vibration tracks the envelope output. The auditory threshold for the BM vibration signal is represented as a low-level additive white noise. Both the envelope and the vibration outputs are available for modeling speech intelligibility and quality.

The reference signal for the intelligibility index is the clean signal without any further processing (no compensation for the hearing loss) which is passed through a model of normal hearing. The reference signal for the quality index is the clean signal combined with NAL-R equalization (Byrne and Dillon, 1986) for the hearing loss and passed through a model of the impaired auditory periphery. The degraded signal for both indices includes the amplification to compensate for the hearing loss and is passed thought the model of the impaired periphery.

The processing for one signal is shown in the block diagram of Fig 2. The auditory model starts with sample rate conversion to 24 kHz, followed by the middle ear filter. The next stage is a linear auditory filter bank, with the filter bandwidths adjusted to reflect the input signal intensity and the increase in filter bandwidth due to outer hair-cell (OHC) damage. Dynamic-range compression is then provided, with the compression controlled by a separate control filter bank. The amount of compression is reduced with increasing OHC damage. Hearing loss due to IHC damage is represented as a subsequent attenuation stage, and IHC firing-rate adaptation is also included in the model. The envelope output in each frequency band comprises the compressed envelope signal after conversion to dB above auditory threshold; the BM vibration signal is centered at the carrier frequency for each band and is modified by the same amplitude modulation as the envelope signal.

Several changes have been made relative to the auditory model used by Kates and Arehart (2010). First is setting the sampling rate to 24 kHz for all signals. Sampling-rate conversion is provided in the model for signals at higher or lower rates. Second is the addition of an intensity adjustment for the auditory filter bandwidth. The shape of the auditory filters depends on the intensity of the input signal as well as on the degree of hearing loss, with the filters becoming broader as the signal intensity increases. Third, the dependence of the outer hair cell (OHC) dynamic range compression with hearing loss has also been adjusted to better reflect the consensus in the literature. Fourth, adaptation of the inner hair-cell (IHC) neural firing rate has been added to the model. Finally, delay compensation has been added to adjust for differences in the auditory filter group delay as a function of band center frequency.

3. Intelligibility and Quality Indices

All three indices are based on signal features computed from the outputs of the auditory model. The equations for the signal features used to construct the indices are given in Appendix A. The indices are monotonic in each of their constituent components. Monotonicity was desired to ensure that valid predictions would be obtained even when the indices are used to predict intelligibility or quality for conditions worse than those present in the training data.

A. HASPI Version 2

The revised HASPI metric is similar to the original in that it is an intrusive metric that compares the output of an auditory model for a reference signal to the model output for a degraded signal. The reference signal is passed through a model of a normal periphery, while the degraded signal is passed through a model of the individual impaired periphery. The peripheral model outputs comprise the envelopes in each auditory filter band, which are combined to form time-varying short-time spectra that are then analyzed using a modulation filterbank. The modulation filter outputs are fitted to the subject intelligibility data using an ensemble of ten neural networks.

The model of the auditory periphery is the one used for the original version of HASPI that is described above. The envelope modulation analysis starts with the dB envelope outputs in each of the thirty-two auditory analysis bands implemented in the peripheral model. The dB envelopes are lowpass filtered with a cutoff frequency of 320 Hz and subsampled. At each time sample at the subsampling rate, the dB envelope values taken across the thirty-two frequency bands represent the log spectrum on an auditory frequency scale. Each short-time spectrum is fit with a set of five basis functions, starting with ½ cycle of a cosine spanning the spectrum from 80 to 8000 Hz and progressing to 2½ cycles spanning the spectrum.

Each of the five cepstral coefficient sequences is passed through a modulation filterbank. The filterbank has ten bands with center frequencies ranging from 2 to 256 Hz. The modulation filtering produces five cepstral coefficient sequences filtered through ten modulation filters. For each of these fifty filtered sequences, the degraded signal is compared to the unprocessed reference signal using a normalized cross-covariance. The cross-correlations are averaged across the basis functions to produce an output vector comprising the averaged covariances for the ten modulation filters.

Neural networks trained using backpropagation were used to map the ten modulation filter outputs to the listener sentence intelligibility scores. One hidden layer comprising four neurons is used, and the output layer comprises a single neuron. The sigmoid activation function is used for all layers. To reduce the possibility of overfitting, the ensemble averaging approach of bootstrap aggregation ("bagging") is used in which the outputs from ten parallel networks are averaged.

B. HASQI Version 2

The HASQI speech quality index compares the cochlear model outputs for a reference signal to the output for a processed (degraded) signal. The reference signal contains amplification to compensate for the hearing loss and is passed through a model of the impaired ear. The degraded signal is also amplified by compensation for the hearing loss and is passed through a model of the impaired auditory periphery. The same set of calculations is used for both normal and impaired hearing, with the only difference being the indicated hearing loss.

HASQI comprises a nonlinear term and a linear term. The nonlinear term combines the cepstral correlation with a vibration correlation term that measures changes in the signal temporal fine structure while ignoring any long-term spectral changes. The linear filtering term compares the long-term spectral representations of the two signals while ignoring the short-term differences in signal modulation and temporal fine structure. The final index output is the combined term, which is the product of the nonlinear term with the linear term.

The nonlinear model is the combination of the cepstral correlation and vibration correlation models. The most accurate combination of the first- and second-order terms and cross products was found to be the product of the square of the cepstral correlation index times the vibration correlation value. The product index is given by:

$$Q_{Nonlin} = c^2 v \quad , \qquad (1)$$

where *c* is the cepstral correlation and *v* is the vibration correlation. The nonlinear term lies between 0 and 1, with 0 indicating no correlation between the degraded and reference signals and 1 indicating perfect correlation.

The linear model is a linear combination of the long-term spectrum and slope standard deviations. A standard deviation of zero gives perfect quality, so the model output is adjusted to start at 1 and is reduced as the standard deviations of the spectrum and slope increase. The model is a MMSE linear regression fit to the combined NH and HI subject ratings for the linear filtered stimuli. The linear model is given by:

$$Q_{Linear} = 1 - 0.579\sigma_1 - 0.421\sigma_2 \quad , \qquad (2)$$

where σ_1 is the standard deviation of the differences in the spectral shape and σ_2 is the standard deviation of the differences in the spectral slope. The linear term is limited to lie between 0 and 1, with 0 indicating poor spectral fidelity and 1 indicating perfect spectral reproduction.

The HASQI version 2 hearing-aid speech quality index is a combination of the index derived for noise and nonlinear processing with the index derived for linear processing. A multiplicative model (Kates and Arehart, 2010) was found to be the most accurate in reproducing the subject ratings for the combined stimuli. The multiplicative model is given by:

$$Q_{Combined} = Q_{Nonlin} \times Q_{Linear} \quad . \qquad (3)$$

In the absence of noise and distortion the multiplicative model is identically the linear index, and in the absence of linear filtering the multiplicative model is identically the noise and nonlinear processing index. The combined index lies between 0 and 1, with 0 indicating the poorest speech quality predicted by the model and 1 indicating perfect reproduction for English sentences.

C. HAAQI version 1

The HAAQI music quality index uses the same general approach as the HASQI speech quality index, but there are differences in formulating HAAQI that reflect the differences between the degraded music and degraded speech quality ratings. Like the speech quality index, the music quality index compares the cochlear model outputs for a reference signal to the output for a processed (degraded) signal. The reference signal contains amplification to compensate for the hearing loss and is passed through a model of the impaired ear. The degraded signal is also amplified by compensation for the hearing loss and is passed through a model of the impaired auditory periphery. The same set of calculations is used for both normal and impaired hearing, with the only difference being the indicated hearing loss.

HAAQI comprises a nonlinear term and a linear term. The nonlinear term combines modulation-filtered cepstral correlation with the same vibration correlation term used for HASQI. The final index output is the combined term, which is a second-order polynomial combination of the nonlinear and linear terms.

The nonlinear model is a combination of the modulation-filtered cepstral correlation and BM vibration terms. For HASQI, as shown in Eq (1), a multiplicative combination was found to be most accurate for predicting speech quality. However, for music quality, a polynomial sum was found to be more accurate. The resultant noise and nonlinear model was found to be:

$$Q_{Nonlin} = 0.246\, v + 0.754\, e^{3} \quad , \qquad (4)$$

where e is the cepstral correlation summed over four modulation filter bands covering 20 to 125 Hz, and v is the vibration correlation.

The linear term for HAAQI is a weighted sum of the spectrum and normalized spectrum standard deviations. The noise and nonlinear term has a maximum value of 1 for perfect signal fidelity, so to be consistent the linear model is adjusted to also start at 1 for no loss of quality and is reduced as the standard deviations of the spectrum and normalized spectrum increase. The linear model, after bootstrap aggregation, is given by:

$$Q_{Linear} = 1 - 0.329\,\sigma_1 - 0.671\,\sigma_3, \tag{5}$$

where σ_1 is the standard deviation of the differences in the spectral shape and σ_3 is the standard deviation of the differences in the normalized spectrum.

The HAAQI version 1 hearing-aid speech quality index is a polynomial combination of the index derived for noise and nonlinear processing with the index derived for linear processing. A second-order regression model (Kates and Arehart, 2015b) was found to be most accurate in reproducing the subject ratings for the combined stimuli. The combined model is given by:

$$Q = 0.336\,Q_{Nonlin} + 0.501\,Q_{Nonlin}^2 + 0.001\,Q_{Linear} + 0.161\,Q_{Linear}^2 \quad . \tag{6}$$

The combined index lies between 0 and 1, with 0 indicating the poorest music quality predicted by the model and 1 indicating perfect reproduction for the three music selections used in creating the index.

4. Using the Indices

The MATLAB code for the function HASPI_v2 is given in Appendix B, the code for HASQI_v2 is given in Appendix C, and the code for HAAQI is given in Appendix D. The headers for the functions indicate the calling arguments and the returned values. These functions in turn call additional functions that implement the cochlear model and extract the signal features. Note that some of the functions called by HASPI_v2, HASQI_v2, and HAAQI_v1 return values that are not used in the index calculations; these values were used in developing the index and it was deemed safer to leave them in the functions rather than trying to rewrite everything to remove the extraneous calculations.

A. HASPI Version 2

The calling sequence for HASPI version 2 is the same as for version 1. The calling argument *x* is the vector containing the reference signal. The vector *x* can be either a column or a row vector; it is internally converted into a row vector by the function. The reference signal should have RMS=1, which corresponds to *Level1* dB SPL, and should have no amplification or compensation for the impaired ear. The reference signal is processed through a model of normal hearing. The argument *fx* is the sampling rate for *x* in Hz. The signal is resampled at 24 kHz internally in the function. If *Level1* is not provided, it defaults to 65 dB SPL.

The calling argument *y* is the vector containing the degraded hearing-aid output signal. The vector *y* can be either a column or a row vector; it is internally converted into a row vector by the function. The amplitude of *y* should be scaled to be RMS=1 prior to the hearing-aid amplification or other signal processing, and compensation for the hearing loss should be provided. The degraded signal thus contains all of the processing to be evaluated and includes compensation for the hearing loss. This signal is processed through a model of the impaired auditory periphery. The argument *fy* is the sampling rate for *y* in Hz. The signal is resampled at 24 kHz internally in the function.

It is best if the sampling rates *fx* and *fy* agree, the lengths of *x* and *y* agree, and that the two signals be temporally aligned. Temporal alignment of the two signals is recommended since exact knowledge of the signal processing system being tested can lead to more accurate alignment than provided by the procedures built into the function. Because of the internal resampling, however, the sampling rates *fx* and *fy* can differ. If the signal lengths differ after resampling, the length of the longer sequence is truncated to match that of the shorter. The function then provides temporal alignment by finding the time delay that maximizes the cross-correlation in each auditory filter band after the signals are passed through the gammatone auditory filters.

The hearing loss is given by row vector *HL*. The loss is specified in dB at the six audiometric frequencies of [250, 500, 1000, 2000, 4000, 6000] Hz. The function applies the hearing loss to the auditory model used for the degraded signal. The normal-hearing model is used for the reference signal.

The function HASPI_v2 returns the computed intelligibility index value and the raw values used in its computation. The intelligibility index is returned in variable *Intel*, which ranges between 0 and 1. The raw values are returned in vector *raw*, which contains the ten cepstral correlation values, averaged over the five basis functions passed through each of the ten modulation filters.

B. HASQI Version 2

The calling argument *x* is the vector containing the reference signal. The vector *x* can be either a column or a row vector; it is internally converted into a row vector by the function. The reference signal should be normalized to have RMS=1 prior to amplification, which corresponds to *Level1* dB SPL. This normalized signal should then be passed through the filter providing linear compensation (e.g. NAL-R) for the hearing loss. The RMS level of the reference signal for impaired hearing will therefore be greater than 1. The reference signal is processed through a model of the impaired auditory periphery. The argument *fx* is the sampling rate for *x* in Hz. The signal is resampled at 24 kHz internally in the function. If *Level1* is not provided, it defaults to 65 dB SPL.

The calling argument *y* is the vector containing the degraded hearing-aid output signal. The vector *y* can be either a column or a row vector; it is internally converted into a row vector by the function. The amplitude of *y* should be scaled to be RMS=1 prior to the hearing-aid amplification or other signal processing, and compensation for the hearing loss should be provided. The degraded signal thus contains all of the processing to be evaluated and includes compensation for the hearing loss. This signal is processed through a model of the impaired auditory periphery. The argument *fy* is the sampling rate for *y* in Hz. The signal is resampled at 24 kHz internally in the function.

It is best if the sampling rates *fx* and *fy* agree, the lengths of *x* and *y* agree, and that the two signals be temporally aligned. Temporal alignment of the two signals is recommended since exact knowledge of the signal processing system being tested can lead to more accurate alignment than the procedures built into the function. Because of the internal resampling, however, the sampling rates *fx* and *fy* can differ. If the signal lengths differ after resampling, the length of the longer sequence is truncated to match that of the shorter. The function then provides temporal alignment by finding the time delay that maximizes the cross-correlation in each auditory filter band after the signals are passed through the gammatone auditory filters.

The hearing loss is given by row vector *HL*. The loss is specified in dB at the six audiometric frequencies of [250, 500, 1000, 2000, 4000, 6000] Hz. The function applies the hearing loss to the auditory model used for both the reference and the degraded signals. The reference signal should include linear amplification to compensate for the hearing loss. If loss compensation has already been applied to the reference, set calling argument *eq*=2. If linear amplification has not been applied, set *eq*=1, which instructs the function to apply NAL-R filtering to the reference signal. For normal hearing *eq* can be set to either 1 or 2. A value of *eq*=2 is preferred since it bypasses the unneeded NAL-R equalization. However, if one sets *eq*=1, the function computes the NAL-R equalization for a hearing loss of 0 dB, which is a default gain of 0 dB at all frequencies.

The function HASQI_v2 returns the combined term that gives the quality prediction, its constituent nonlinear and linear terms, and the raw values used in its computation. The HASQI quality index is returned in variable *Combined*, and the nonlinear and linear terms are returned in variables *Nonlin* and *Linear*, respectively. The nonlinear term is computed using Eq (1), the linear term is computed using Eq (2), and the combined quality prediction is computed using Eq (3). The raw values are returned in vector *raw*, which contains the cepstral correlation, the vibration correlation, the standard deviation of the spectral differences, and the standard deviation of the differences in the spectral slopes.

C. HAAQI Version 1

The calling sequence for HAAQI is identical to that for HASQI, and the function is used in exactly the same way. The function HAAQI_v1 returns the combined term that gives the quality prediction, its constituent nonlinear and linear terms, and the raw values used in its computation. The HAAQI quality index is returned in variable *Combined*, and the nonlinear and linear terms are returned in variables *Nonlin* and *Linear*, respectively. The nonlinear term is computed using Eq (5), the linear term is computed using Eq (5), and the combined quality prediction is computed using Eq (6). The raw values are returned in vector *raw*, which contains the modulation-filtered cepstral correlation, the vibration correlation, the standard deviation of the spectral differences, and the standard deviation of the normalized spectral differences.

D. Level Calibration

The auditory model used for HASPI, HASQI, and HAAQI incorporates the auditory threshold, and the model processes the signals based on their intensity relative to threshold. Signal level calibration is therefore critical, since the model needs to know where the reference and processed signals lie relative to auditory threshold. The recommended procedure is that the unprocessed unamplified reference signal be set to have its RMS intensity equal to 1. The parameter *Level1* used in calling HASPI and HASQI can then be set to level used for speech or music signal presentation, typically 65 dB SPL; the default if *Level1* is not provided is 65 dB SPL. If NAL-R or equivalent amplification is provided to compensate for the hearing loss, or if hearing-aid processing is being used, the processing should be applied after the signal RMS level has been set to 1.

E. Example

Two sound files are included with this memorandum. The file sig_clean.wav contains two concatenated sentences from the HINT corpus spoken by a male talker. This signal is the input to a simulated hearing aid. The file sig_out is the time-aligned output from the simulation. The simulation includes additive noise, peak-clipping distortion, and high-pass and low-pass filtering of the signal. The simulation includes a frequency-warped dynamic-range compressor filter bank, and the input signal has been passed through the all-pass filter cascade to match its group delay to that of the output signal. The sampling rate for both signals is 22050 Hz.

To check the implementation of the metrics, read sig_clean.wav into MATLAB using audioread and assign it to vector *x* and its sampling rate to *fx*. Read in sig_out.wav and assign it to vector *y* and its sampling rate to *fy*. Normalize each signal so that it has a RMS amplitude of 1. Set *HL*=[0, 0, 0, 0, 0, 0] for normal hearing, and set *Level1*=65. For HASQI_v2, set *eq*=2 since NAL-R equalization is not needed for normal hearing. The estimated intelligibility returned by HASPI_v2 is *Intel*= 0.6427, which is higher than the value 0f 0.463 returned by version 1 due to the change in the metric calculation procedure. The quality index HASQI_v2 returns *Combined*=0.072 as the quality prediction, with the its constituent terms *Nonlin*=.110 and *Linear*=.657.

5. Comparisons among Indices

The predictions produced by the new and old speech indices are compared in Fig 3. The stimulus was the concatenation of 20 IEEE sentences chosen at random and using both female and male talkers. The psychometric function for HASPI v2 is similar to that for v1 at SNRs of 5 dB and above but is steeper in the vicinity of the 50-percent correct point. Version 2 is thus more sensitive to the presence of additive noise over the range of -5 to 5 dB SNR than the previous version, and would be expected to give lower intelligibility predictions for many situations of additive LTASS noise.

The differences between the HASQI version 1 and HASQI version 2 quality predictions are small for the additive noise. Thus the quality predictions in the figure are quite similar, even though the predictions for other types of signal degradations such as frequency compression and noise vocoder output can differ by much larger amounts (Kates and Arehart, 2014b).

The differences between HAAQI and HASQI for a segment of jazz music are illustrated in Fig 4 as the SNR is varied. The music has been combined with the same stationary speech-shaped noise as used for Fig 3. HASQI applied to the music selection gives a curve similar to that for the speech. However, HASQI gives a higher quality rating than HAAQI for SNRs between 0 and 30 dB and lower ratings for SNRs above 30 dB. HAAQI represents a more accurate fit to the music quality ratings than does HASQI, so the differences in the two curves indicate how quality changes with the stimuli being evaluated and illustrates the potential error in applying an index to stimuli for which it has not been validated.

6. Limitations

HASPI version 2, HASQI version 2, and HAAQI version 1 provide accurate predictions for an average listener with normal hearing or having the indicated audiogram. The predictions do not account for individual variation in auditory abilities or cognitive function that cannot be explained by the audiogram. There will therefore be substantial inter-subject variability when the index predictions are applied to an individual listener. The index predictions thus provide the starting point for an algorithm evaluation or hearing-aid fitting, but individual fine-tuning will be needed in any clinical application.

A potential limitation of the indices concerns temporal alignment. The main application of HASPI, HASQI, and HAAPI is hearing aids rather than telecommunications systems. Thus aspects of digital speech encoding and transmission, such as gradual changes in the timebase over the course of a sentence, packet loss, and potential rapid realignment of the timebase following pauses, are not included in the model or in the provided temporal alignment. The cepstral correlation and short-time coherence calculations are not expected to be affected by small timing changes, but the new indices need to be evaluated for their sensitivity to timing effects and the accuracy of its predictions for digital telecommunications systems.

A second potential limitation of the new indices is that they were derived using data for monaural headphone listening. Further research is needed to deal with the effects of binaural listening and the impact of room reverberation in the binaural scenario.

A further consideration is that the models are based on sentences in English or on a limited number of music selections. Intelligibility for English words, rather than sentences, would require different weights applied to the cepstral correlation and auditory coherence values prior to the logistic transform. Extending the model to other languages is also an issue; the relative importance of the envelope and the temporal fine structure for predicting intelligibility and quality is expected to differ for tonal languages in comparison with English. Applying HAAQI to other genres of music, such as rock, may also lead to inaccurate results since the amount of nonlinear distortion that is considered acceptable may differ from the classical and jazz selections used in forming the index.

7. References

Byrne, D., and Dillon, H. (**1986**). “The national acoustics laboratories’ (NAL) new procedure for selecting gain and frequency response of a hearing aid,” Ear Hear. **7**, 257-265.

Kates, J.M. (**2013**). “An auditory model for intelligibility and quality predictions,” Proc. Mtgs. Acoust. (POMA) **19**, 050184: Acoust. Soc. Am. 165th Meeting, Montreal, June 2-7, 2013.

Kates, J.M., and Arehart, K.H. (**2005**). “Coherence and the Speech Intelligibility Index,” J. Acoust. Soc. Am. **117**, 2224-2237.

Kates, J.M., and Arehart, K.H. (**2010**). “The hearing aid speech quality index (HASQI),” J. Audio Eng. Soc. **58**, 363-381.

Kates, J.M., and Arehart, K.H. (**2014a**). “The hearing aid speech perception index (HASPI),” Speech Comm. **65**, 75-93.

Kates, J.M., and Arehart, K.H. (**2014b**). “The hearing-aid speech quality index (HASQI) version 2,” J. Audio Eng. Soc. **62**, 99-117.

Kates, J.M. and Arehart, K.H. (**2015a**), “Comparing the information conveyed by envelope modulation for speech intelligibility, speech quality, and music quality”, J. Acoust. Soc. Am. **138**, 2470-2482.

Kates, J.M., and Arehart, K.H. (**2015b**). “The hearing aid audio quality index (HAAQI),” IEEE Trans. Audio Speech and Lang. Proc. **24**, 354-365.

Kates, J.M., and Arehart, K.H. (**2020**), “The hearing-aid speech perception index (HASPI) version 2,” submitted for publication.

Appendix A: Signal Features

The cepstral correlation and auditory coherence features are used in the HASPI calculation. The cepstral correlation, vibration correlation, and spectral shape features are used in the HASQI calculation.

A. Cepstral Correlation

The envelope samples output by the auditory model, when taken across frequency at a given time slot, constitute a short-time log magnitude spectrum on an auditory frequency scale. The inverse Fourier transform of this log spectrum produces a set of coefficients that are similar to the mel cepstrum. In the model, only a small number of cepstrum coefficients are needed, so the cepstrum computation is performed in the frequency domain by fitting the auditory model envelope outputs with a set of half-cosine basis functions. These basis functions are very similar to the principal components for the short-time spectra of speech. The basis functions are given by:

$$b_j(k) = \cos[(j-1)\pi k/(K-1)] \quad , \qquad \text{(A1)}$$

where j is the basis function number and k is the gammatone filter index for frequency bands 0 though K-1 for K=32.

Let $e_k(m)$ denote the sequence of smoothed sub-sampled envelope samples in frequency band k for the reference signal, and let $d_k(m)$ be the envelope samples for the degraded signal. The envelope smoothing is provided by 16-ms von Hann windows having 50-percent overlap, giving a lowpass filter cutoff frequency of 62.5 Hz and a smoothed envelope sampling rate of 125 Hz. The reference-signal cepstral sequence $p_j(m)$ and the degraded-signal sequence $q_j(m)$ are then given by:

$$p_j(m) = \sum_{k=0}^{K-1} b_j(k) e_k(m)$$
$$q_j(m) = \sum_{k=0}^{K-1} b_j(k) d_k(m) \quad . \qquad \text{(A2)}$$

The cepstrum correlation is computed by taking the cross-covariance of the cepstral sequences for the reference and degraded signals. The sequence values corresponding to pauses in the speech are removed from cepstral sequences $p_j(m)$ and $q_j(m)$, and the mean value is subtracted from each edited sequence to yield the zero-mean edited sequences $\hat{p}_j(m)$ and $\hat{q}_j(m)$. The pauses were deleted by converting the envelopes in each band to linear values to give numbers related to specific loudness (Kates, 2013), summing the values across frequency, and then converting the sum back to dB. Segments having an intensity less than 2.5 dB re:threshold were removed from the correlation calculation. The normalized covariance is then given by:

$$r(j) = \frac{\sum_{m \in Speech} \hat{p}_j(m)\hat{q}_j(m)}{\left[\sum_{m \in Speech} \hat{p}_j^2(m)\right]^{1/2} \left[\sum_{m \in Speech} \hat{q}_j^2(m)\right]^{1/2}} \quad . \qquad \text{(A3)}$$

The average cepstrum correlation given by the average of the normalized covariance values *r*(2) though *r*(6):

$$c = \frac{1}{5}\sum_{j=2}^{6} r(j) \quad . \qquad \text{(A4)}$$

B. HASPI Modulation Filtered Cepstral Correlation

HASPI v2 uses the cepstral sequences passed through a modulation filterbank. The envelope of the signal in each auditory filter band is extracted via the complex demodulation used for the gammatone filters. The envelope in each frequency band is converted to dB within the auditory model and lowpass filtered at 320 Hz using a raised-cosine window to preclude the possibility of negative envelope samples. The lowpass-filters envelopes are then subsampled at 2560 Hz and the five signals $p_j(m)$ and $q_j(m)$ are computed using Eq (A2) at the 2560-Hz sub-sampling rate.

Each of the five basis function sequences is passed through a modulation filterbank comprising ten filters having center frequencies from 2 to 256 Hz. The filtering operation uses complex modulation to rotate the signal down to baseband, followed by a raised-cosine linear-phase FIR lowpass filter and complex demodulation back to the band carrier frequency. The filter responses were adjusted to have a Q of 1.5.

For each modulation filter output and basis function sequence, the time-frequency envelope pattern of the degraded signal being evaluated is compared to the envelope of the unprocessed reference signal using normalized cross-covariance, producing a value between 0 and 1. Let $u_{j,n}(m)$ be the basis function signal $p_j(m)$ passed through modulation filter n, and let $v_{j,n}(m)$ be the basis function signal $q_j(m)$ passed through modulation filter n. The cross-covariance between the degraded and reference signals is then given by:

$$r(j,n) = \frac{\sum_{m \in Speech} u_{j,n}(m)v_{j,n}(m)}{\left[\sum_{m \in Speech} u_{j,n}^2(m)\right]^{1/2} \left[\sum_{m \in Speech} v_{j,n}^2(m)\right]^{1/2}} \quad . \qquad \text{(A5)}$$

A total of fifty cross-covariance calculations are produced for the five basis function sequences and ten modulation filters. The cross-covariance values are averaged over the five basis functions, producing ten lowpass filter values that form the inputs to the neural networks.

An ensemble of ten neural networks was used to map the modulation filter outputs to the listener intelligibility scores. The scores were expressed as proportion sentences correct, giving values over [0,1]. The neural-network approach was chosen for its ability to approximate an arbitrary nonlinear function and for its ability to model of potential interactions between the input variables. The ten inputs to each neural network were the averaged covariances for the ten modulation filters produced by the envelope modulation analysis. One hidden layer comprising four neurons was used, and the output layer comprised a single neuron. The sigmoid activation function was used for all layers; the sigmoid function applied to the output ensures that it is bounded between 0 and 1 to match the range of the sentence-correct scores. The neural networks were trained using basic backpropagation with a mean-squared error loss function.

C. HAAQI Modulation Filtered Cepstral Correlation

HAAQI also uses the cepstral sequences passed through a modulation filter bank. The HAAQI filterbank differs from the one used for HASPI v2. The signal envelope in each auditory frequency is smoothed using 8-ms von Hann windows having 50% overlap, giving a lowpass filter cutoff frequency of 125 Hz and a smoothed envelope sampling rate of 250 Hz.

The five signals $p_j(m)$ and $q_j(m)$ are computed using Eq (A2) at the 250-Hz sub-sampling rate. Each signal is passed through a modulation filterbank comprising eight filters covering 0 to 125 Hz, implemented using 128-sample linear-phase finite impulse-response (FIR) filters at the 250-Hz sub-sampling rate. The leading and trailing filter transients are removed.

For each modulation filter output and basis function sequence, the time-frequency envelope pattern of the degraded signal being evaluated is compared to the envelope of the unprocessed reference signal using normalized cross-covariance, producing a value between 0 and 1. The filtered basis functions are cross-correlated using Eq (A5). The results of Kates and Arehart (2015a) indicate that the four highest modulation frequencies convey the greatest amount of music quality information, so the cepstral correlation term used in HAAQI is the average of the cross-covariances for the four highest modulation frequency bands, covering 20 through 125 Hz. The cepstral correlation, averaged over basis functions 2-6 and modulation frequency bands 5-8, is then:

$$e=\frac{1}{5}\frac{1}{4}\sum_{j=2}^{6}\sum_{n=5}^{8} r(j,n) \quad . \qquad \text{(A6)}$$

D. Auditory Coherence

The basilar membrane output of the auditory model was divided into 16-ms segments having a 50-percent overlap, with each segment multiplied by a von Hann window. The intensity and short-time coherence was computed for each segment in each auditory frequency band. The intensity of the vibration output from the auditory model was in dB SL. The intensity in each segment of the reference signal was converted from log to linear amplitude, and the segment intensities summed across frequencies to form a broadband intensity signal. The segments of the reference signal that correspond to silent intervals were identified, and the silent segments in the reference and degraded signals were discarded. A cumulative histogram of the intensities of the remaining segments was then created, with segments assigned to either the lowest third, middle third, or upper third of the histogram.

The short-time coherence values for the low-level, mid-level, and high-level segments were then averaged across time and frequency to produce the low-, mid-, and high-level auditory coherence values. Let $x_k(m,n)$ be the BM vibration for the reference signal and $y_k(m,n)$ be the BM vibration for the degraded signal in frequency band k and segment m. The signals after being windowed and converted to zero-mean are given by $\hat{x}_k(m,n)$ and $\hat{y}_k(m,n)$. The normalized cross correlation for segment *m* in frequency band *k* is given by:

$$z(m,k) = \underset{\tau}{Max}\left\{ \frac{\sum_n \hat{x}_k(m,n)\hat{y}_k(m,n+\tau)}{\left[\sum_n \hat{x}_k^2(m,n)\right]^{1/2}\left[\sum_n \hat{y}_k^2(m,n)\right]^{1/2}} \right\}, \qquad \text{(A7)}$$

where the delay *τ* is chosen over the range of -1 to 1 ms to yield the maximum value of the cross-correlation. The values of *z*(*m*,*k*) for each of the three intensity levels were averaged to produce the associated auditory coherence values, with a_{Low} being the low-level auditory coherence value, a_{Mid} the mid-level value, and a_{High} the high-level value.

E. Vibration Correlation

The vibration correlation calculation starts with the normalized cross-correlation values *z*(*m*,*k*) given by Eq (A5). Each normalized cross-correlation value *z*(*m*,*k*) is multiplied by a frequency-dependent weight *w*(*m*,*k*) that is set to 0 if the segment of the reference speech lies below threshold and is set to the IHC synchronization index for segments above threshold. The synchronization index gives the degree to which the neural firing pattern is synchronized to the temporal fluctuations of the signal within the band, and decreases with increasing frequency. A fifth-order lowpass filter with a cutoff frequency of 3.5 kHz was found to yield the most-accurate quality predictions. The weighted cross-correlations are summed across segment and frequency band and are divided by the sum of the weights to give the vibration correlation:

$$v = \frac{\sum_k \sum_m w(m,k) z(m,k)}{\sum_k \sum_m w(m,k)} .$$

(A8)

F. Spectral Shape

The linear index for predicting speech quality is based on changes to the long-term signal spectrum. Linear filters and spectral modifications all affect the long-term spectrum of the signal but have essentially no effect on the envelope correlations between the reference and filtered signals that are measured by the nonlinear model. The linear model for measuring the long-term spectral effects is the same approach as used by Kates and Arehart (2010).

The average output of the system is the signal levels in each analysis band converted to dB re:threshold. Prior to the calculations the levels are converted back to linear amplitude, and the input and output spectra are normalized to both have RMS values of 1 when summed across the complete set of auditory bands. The level normalization removes the signal amplitude as a factor in the linear model, leaving only the spectral differences as factors.

Let $\left|\hat{X}(k)\right|$ be the normalized input linear signal spectrum, and $\left|\hat{Y}(k)\right|$ be the normalized output signal spectrum. The difference in the spectra is given by:

$$d_1(k) = \left|\hat{Y}(k)\right| - \left|\hat{X}(k)\right| \text{ , } 0 \le k \le K\text{-}1 \quad , \qquad \text{(A9)}$$

and the difference in the spectral slopes is given by:

$$d_2(k) = \left|\left|\hat{Y}(k)\right| - \left|\hat{Y}(k-1)\right|\right| - \left|\left|\hat{X}(k)\right| - \left|\hat{X}(k-1)\right|\right| \text{ , } 1 \le k \le K\text{-}1 \quad . \qquad \text{(A10)}$$

The standard deviation of the spectral difference is then:

$$\sigma_1 = g_1 \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \left[d_1(k) - \bar{d}_1 \right]^2 \right\}^{1/2} \quad , \qquad \text{(A11)}$$

where g_1 is a scaling factor empirically set to 0.4, and the standard deviation of the slope difference is given by:

$$\sigma_2 = \left\{ \frac{1}{K-1} \sum_{k=1}^{K-1} \left[d_2(k) - \bar{d}_2 \right]^2 \right\}^{1/2} \quad , \qquad \text{(A12)}$$

where the overbar in both equations denotes the average over the frequency bands. Both standard deviations have a minimum value of 0, and were limited to have a maximum value of 1.

HAAQI uses the spectral difference given by Eq (A11) combined with a second form of spectral difference which is normalized by the intensities in each frequency band:

$$d_3(k) = \frac{\hat{X}(k) - \hat{Y}(k)}{\hat{X}(k) + \hat{Y}(k)} \quad , 0 \le k \le K\text{-}1 . \qquad \text{(A13)}$$

The standard deviation of the normalized spectral differences is given by:

$$\sigma_3 = g_3 \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \left[d_3(k) - \bar{d}_3 \right]^2 \right\}^{1/2} \quad , \qquad \text{(A14)}$$

where the scaling factor g_3 is set to 0.04 . This standard deviation also has a minimum value of 0, indicating no spectral modification, and the maximum value is limited to 1.

Appendix B: HASPI_v2 .m

```
function [Intel,raw] = HASPI_v2(x,fx,y,fy,HL,Level1)
% Function to compute the HASPI intelligibility index using the
% auditory model followed by computing the envelope cepstral
% correlation and BM vibration high-level covariance. The reference
% signal presentation level for NH listeners is assumed to be 65 dB
% SPL. The same model is used for both normal and impaired hearing.
This
% version of HASPI uses a modulation filterbank followed by an
ensemble of
% neural networks to compute the estimated intelligibility.
%
% Calling arguments:
% x             Clear input reference speech signal with no noise or
distortion.
%           If a hearing loss is specified, no amplification should be
provided.
% fx        Sampling rate in Hz for signal x
% y             Output signal with noise, distortion, HA gain, and/or
processing.
% fy        Sampling rate in Hz for signal y.
% HL       (1,6) vector of hearing loss at the 6 audiometric
frequencies
%                 [250, 500, 1000, 2000, 4000, 6000] Hz.
% Level1    Optional input specifying level in dB SPL that corresponds
to a
%           signal RMS = 1. Default is 65 dB SPL if argument not
provided.
%
% Returned values:
% Intel     Intelligibility estimated by passing the cepstral
coefficients
%           through a modulation filterbank followed by an ensemble of
%           neural networks.
% raw       vector of 10 cep corr modulation filterbank outputs,
averaged
%           over basis funct 2-6.
%
% James M. Kates, 5 August 2013.

% Set the RMS reference level
if nargin < 6
    Level1=65;
end
```

```
% Auditory model for intelligibility
% Reference is no processing, normal hearing
itype=0; %Intelligibility model
[xenv,xBM,yenv,yBM,xSL,ySL,fsamp]=...
    eb_EarModel(x,fx,y,fy,HL,itype,Level1);

% ---------------------------------------
% Envelope modulation features
% LP filter and subsample the envelopes
fLP=320; %Envelope LP cutoff frequency, Hz
fsub=8*fLP; %Subsample to span 2 octaves above the cutoff frequency
[xLP,yLP]=ebm_EnvFilt(xenv,yenv,fLP,fsub,fsamp);

% Compute the cepstral coefficients as a function of subsampled time
nbasis=6; %Use 6 basis functions
thr=2.5; %Silence threshold in dB SL
dither=0.1; %Dither in dB RMS to add to envelope signals
[xcep,ycep]=ebm_CepCoef(xLP,yLP,thr,dither,nbasis);

% Cepstral coefficients filtered at each modulation rate
% Band center frequencies [2, 6, 10, 16, 25, 40, 64, 100, 160, 256] Hz
% Band edges [0, 4, 8, 12.5, 20.5, 30.5, 52.4, 78.1, 128, 200, 328] Hz
[xmod,ymod,cfmod]=ebm_ModFilt(xcep,ycep,fsub);

% Cross-correlations between the cepstral coefficients for the
degraded and
% ref signals at each modulation rate, averaged over basis functions
2-6
% aveCM:  cep corr modulation filterbank outputs, ave over basis funct
2-6
aveCM=ebm_ModCorr(xmod,ymod);

% ---------------------------------------
% Intelligibility prediction
% Get the neural network parameters and the weights for an ensemble of
% 10 networks
[NNparam,Whid,Wout,b]=ebm_GetNeuralNet;

% Average the neural network outputs for the modulation filterbank
values
model=NNfeedfwdEns(aveCM,NNparam,Whid,Wout);
model=model/b;

% Return the intelligibility estimate and raw modulation filter
outputs
Intel=model;
raw=aveCM;
end
```

Appendix C: HASQI_v2.m

```
function [Combined,Nonlin,Linear,raw] =
HASQI_v2(x,fx,y,fy,HL,eq,Level1)
% Function to compute the HASQI version 2 quality index using the
% auditory model followed by computing the envelope cepstral
% correlation and BM vibration average short-time coherence signals.
% The reference signal presentation level for NH listeners is assumed
% to be 65 dB SPL. The same model is used for both normal and
% impaired hearing.
%
% Calling arguments:
% x        Clear input reference speech signal with no noise or
distortion.
%           If a hearing loss is specified, NAL-R equalization is
optional
% fx        Sampling rate in Hz for signal x
% y        Output signal with noise, distortion, HA gain, and/or
processing.
% fy        Sampling rate in Hz for signal y.
% HL       (1,6) vector of hearing loss at the 6 audiometric
frequencies
%                 [250, 500, 1000, 2000, 4000, 6000] Hz.
% eq        Flag to provide equalization for the hearing loss to
signal x:
%             1 = no EQ has been provided, the function will add NAL-R
%             2 = NAL-R EQ has already been added to the reference
signal
% Level1    Optional input specifying level in dB SPL that corresponds
to a
%           signal RMS = 1. Default is 65 dB SPL if argument not
provided.
%
% Returned values:
% Combined  Quality estimate is the product of the nonlinear and
linear terms
% Nonlin    Nonlinear quality component = (cepstral corr)^2 x seg BM
coherence
% Linear    Linear quality component = std of spectrum and spectrum
slope
% raw       Vector of raw values = [CepCorr, BMsync5, Dloud, Dslope]
%
% James M. Kates, 5 August 2013.

% Set the RMS reference level
if nargin < 7
    Level1=65;
end

% Auditory model for quality
% Reference is no processing or NAL-R, impaired hearing
[xenv,xBM,yenv,yBM,xSL,ySL,fsamp]=...
    eb_EarModel(x,fx,y,fy,HL,eq,Level1);

% ----------------------------------------
% Envelope and long-term average spectral features
% Smooth the envelope outputs: 125 Hz sub-sampling rate
segsize=16; %Averaging segment size in msec
xdB=eb_EnvSmooth(xenv,segsize,fsamp);
ydB=eb_EnvSmooth(yenv,segsize,fsamp);

% Mel cepstrum correlation using smoothed envelopes
% m1=ave of coefficients 2-6
% xy=vector of coefficients 1-6
thr=2.5; %Silence threshold: sum across bands, dB above aud threshold
addnoise=0.0; %Additive noise in dB SL to condition cross-covariances
[CepCorr,xy]=eb_melcor(xdB,ydB,thr,addnoise);

% Linear changes in the log-term spectra
% dloud  vector: [sum abs diff, std dev diff, max diff] spectra
% dnorm  vector: [sum abs diff, std dev diff, max diff] norm spectra
% dslope vector: [sum abs diff, std dev diff, max diff] slope
[dloud,dnorm,dslope]=eb_SpectDiff(xSL,ySL);

% ----------------------------------------
% Temporal fine structure correlation measurements
% Compute the time-frequency segment covariances
segcov=16; %Segment size for the covariance calculation
[sigcov,sigMSx,sigMSy]=eb_BMcovary(xBM,yBM,segcov,fsamp);

% Average signal segment cross-covariance
% avecov=weighted ave of cross-covariances, using only data above
threshold
% syncov=ave cross-covariance with added IHC loss of synchronization
at HF
thr=2.5; %Threshold in dB SL for including time-freq tile
[avecov,syncov]=eb_AveCovary2(sigcov,sigMSx,thr);
BMsync5=syncov(5); %Ave segment coherence with IHC loss of sync

% Extract and normalize the spectral features
% Dloud:std
d=dloud(2); %Loudness difference std
d=d/2.5; %Scale the value
d=1.0 - d; %1=perfect, 0=bad
d=min(d,1);
d=max(d,0);
Dloud=d;

% Dslope:std
d=dslope(2); %Slope difference std
```

```
d=1.0 - d;
d=min(d,1);
d=max(d,0);
Dslope=d;

% ------------------------------------
% Construct the models
% Nonlinear model
Nonlin=(CepCorr^2)*BMsync5; %Combined envelope and temporal fine
structure

% Linear model
Linear=0.579*Dloud + 0.421*Dslope; %Linear fit

% Combined model
Combined=Nonlin*Linear; %Product of nonlinear x linear

% Raw data
raw=[CepCorr,BMsync5,Dloud,Dslope];

end
```

Appendix D: HAAQI_v1.m

```
function [Combined,Nonlin,Linear,raw] =
HAAQI_v1(x,fx,y,fy,HL,eq,Level1)
% Function to compute the HAAQI music quality index using the
% auditory model followed by computing the envelope cepstral
% correlation and BM vibration average short-time coherence signals.
% The reference signal presentation level for NH listeners is assumed
% to be 65 dB SPL. The same model is used for both normal and
% impaired hearing.
%
% Calling arguments:
% x        Clear input reference speech signal with no noise or
distortion.
%           If a hearing loss is specified, NAL-R equalization is
optional
% fx        Sampling rate in Hz for signal x
% y        Output signal with noise, distortion, HA gain, and/or
processing.
% fy        Sampling rate in Hz for signal y.
% HL       (1,6) vector of hearing loss at the 6 audiometric
frequencies
%                 [250, 500, 1000, 2000, 4000, 6000] Hz.
% eq        Flag to provide equalization for the hearing loss to
signal x:
%             1 = no EQ has been provided, the function will add NAL-R
%            2 = NAL-R EQ has already been added to the reference
signal
% Level1    Optional input specifying level in dB SPL that corresponds
to a
%           signal RMS = 1. Default is 65 dB SPL if argument not
provided.
%
% Returned values:
% Combined  Quality is the polynomial sum of the nonlin and linear
terms
% Nonlin    Nonlinear quality component = .245(BMsync5)
+ .755(CepHigh)^3
% Linear    Linear quality component = std of spectrum and norm
spectrum
% raw       Vector of raw values = [CepHigh, BMsync5, Dloud, Dnorm]
%
% James M. Kates, 5 August 2013 (HASQI_v2).
% Version for HAAQI_v1, 19 Feb 2015.

% Set the RMS reference level
if nargin < 7
    Level1=65;
end

% Auditory model for quality
% Reference is no processing or NAL-R, impaired hearing
[xenv,xBM,yenv,yBM,xSL,ySL,fsamp]=...
    eb_EarModel(x,fx,y,fy,HL,eq,Level1);

% ---------------------------------------
% Envelope and long-term average spectral features
% Smooth the envelope outputs: 250 Hz sub-sampling rate
segsize=8; %Averaging segment size in msec
xdB=eb_EnvSmooth(xenv,segsize,fsamp);
ydB=eb_EnvSmooth(yenv,segsize,fsamp);

% Mel cepstrum correlation after passing through modulation filterbank
thr=2.5; %Silence threshold: sum across bands, dB above aud threshold
addnoise=0.0; %Additive noise in dB SL to condition cross-covariances
[CepAve,CepLow,CepHigh,CepModVector]= ...
    eb_melcor9(xdB,ydB,thr,addnoise,segsize); %8 modulation freq bands

% Linear changes in the long-term spectra
% dloud  vector: [sum abs diff, std dev diff, max diff] spectra
% dnorm  vector: [sum abs diff, std dev diff, max diff] norm spectra
% dslope vector: [sum abs diff, std dev diff, max diff] slope
[dloud,dnorm,dslope]=eb_SpectDiff(xSL,ySL);

% ---------------------------------------
% Temporal fine structure (TFS) correlation measurements
% Compute the time-frequency segment covariances
segcov=16; %Segment size for the covariance calculation
[sigcov,sigMSx,sigMSy]=eb_BMcovary(xBM,yBM,segcov,fsamp);

% Average signal segment cross-covariance
% avecov=weighted ave of cross-covariances, using only data above threshold
% syncov=ave cross-covariance with added IHC loss of synchronization at HF
thr=2.5; %Threshold in dB SL for including time-freq tile
[avecov,syncov]=eb_AveCovary2(sigcov,sigMSx,thr);
BMsync5=syncov(5); %Ave segment coherence with IHC loss of sync

% Extract and normalize the spectral features
% Dloud:std
d=dloud(2); %Loudness difference std
d=d/2.5; %Scale the value
d=1.0 - d; %1=perfect, 0=bad
d=min(d,1);
d=max(d,0);
Dloud=d;

% Dnorm:std
d=dnorm(2); %Slope difference std
d=d/25; %Scale the value
d=1.0 - d; %1=perfect, 0=bad
d=min(d,1);
d=max(d,0);
Dnorm=d;

% ----------------------------------------
% Construct the models
% Nonlinear model
Nonlin=0.754*(CepHigh^3) + 0.246*BMsync5; %Combined envelope and TFS

% Linear model
Linear=0.329*Dloud + 0.671*Dnorm; %Linear fit

% Combined model
Combined=0.336*Nonlin + 0.001*Linear + 0.501*(Nonlin^2) + 0.161*(Linear^2); %Polynomial sum

% Raw data
raw=[CepHigh,BMsync5,Dloud,Dnorm];

end
```

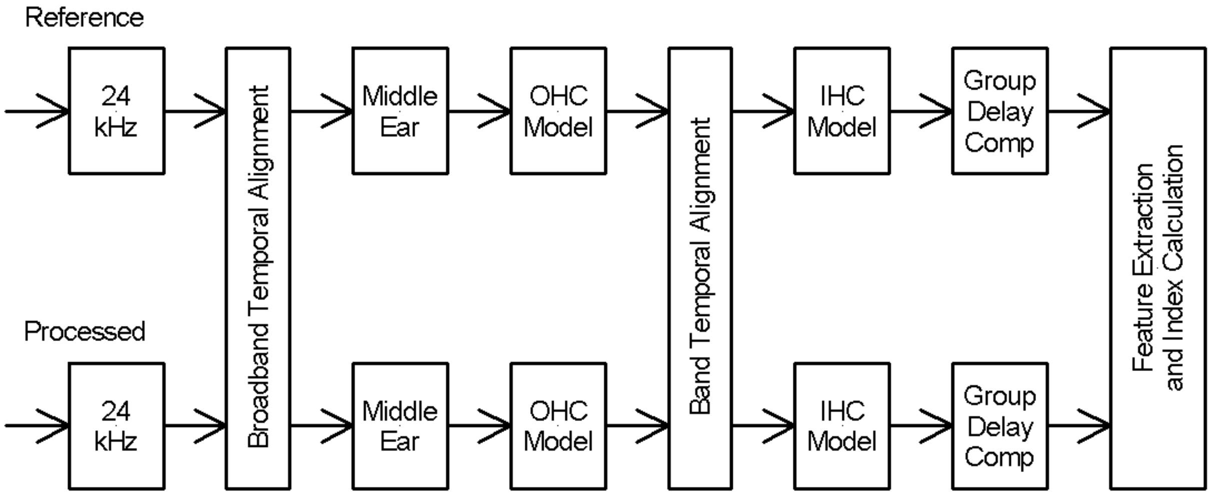

Fig 1. Block diagram of the reference and processed signal comparison.

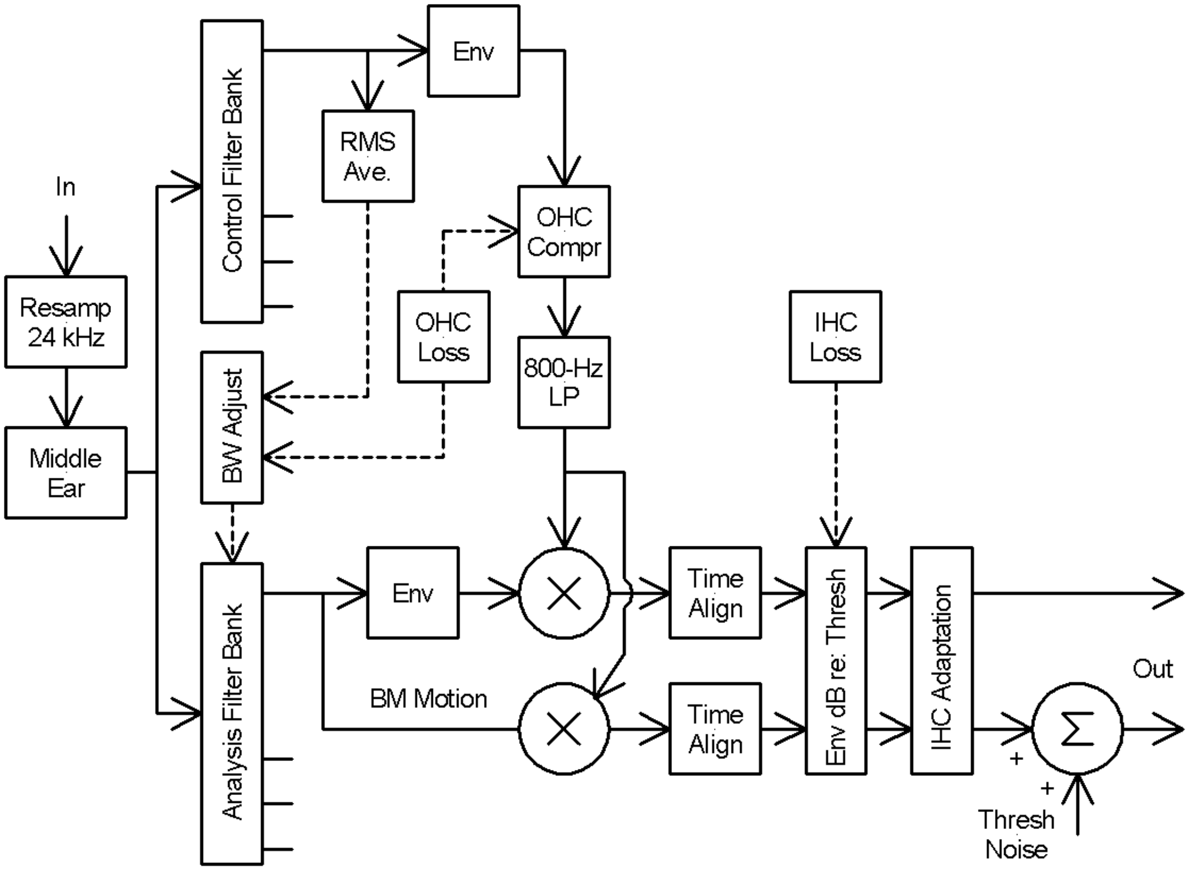

Fig. 2. Block diagram of the auditory model used to extract the signals in each frequency band.

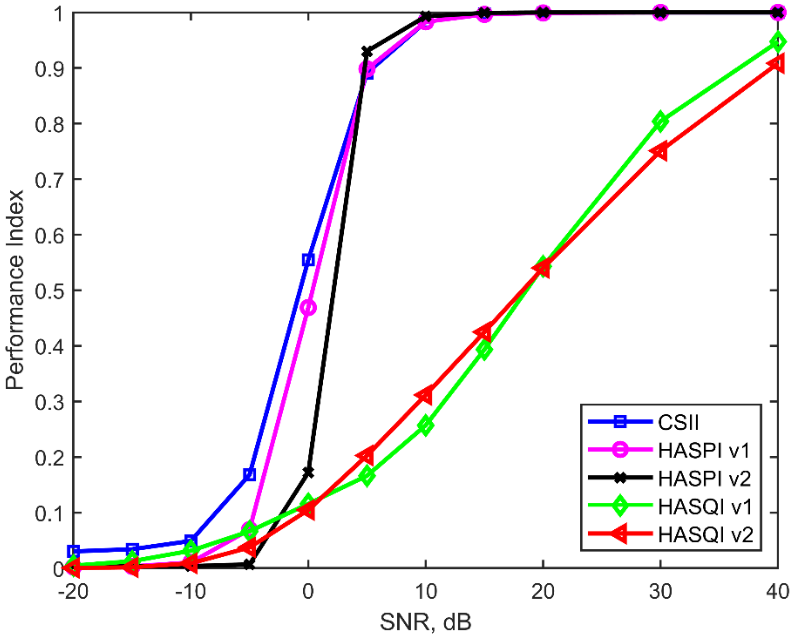

Fig 3. Comparison of the new and old indices for 20 IEEE sentences combined with additive speech-shaped noise.

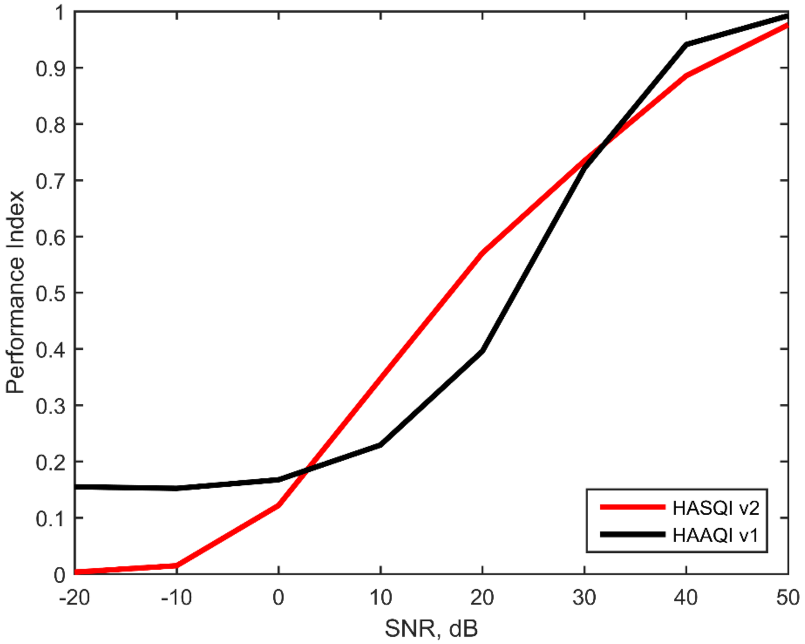

Fig 4. Comparison of HAAQI and HASQI version 2 for a selection of jazz music with additive speech-shaped noise.