Speech Communication 131 (2021) 35–46

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

The Hearing-Aid Speech Perception Index (HASPI) Version 2

James M. Kates*, Kathryn H. Arehart

Department of Speech Language and Hearing Sciences, University of Colorado, Boulder, CO 80309, USA

ARTICLE INFO

Keywords:
Speech intelligibility
Intelligibility index
Auditory model
Auditory amplitude modulation detection
Hearing loss
Hearing aids

ABSTRACT

This paper presents a revised version of the Hearing-Aid Speech Perception Index (HASPI). The index is based on a model of the auditory periphery that incorporates changes due to hearing loss and is valid for both normal-hearing and hearing-impaired listeners. It is an intrusive metric that compares the time-frequency envelope and temporal fine structure (TFS) of a degraded signal to an unprocessed reference. The first modification to HASPI is an extension to the range of envelope modulation rates considered in the metric. HASPI applies a lowpass filter to the time-frequency envelope modulation, and in the new version this single filter is replaced by a modulation filterbank. The temporal fine structure (TFS) analysis in the original version of HASPI is replaced by the filterbank outputs at higher modulation rates that represent auditory roughness and periodicity. The second modification is replacing the parametric model combining envelope and TFS measurements used in the original version with an ensemble of neural networks. The improved version of HASPI is compared to the original version for datasets from five experiments that encompass noise and nonlinear distortion, frequency compression, ideal binary mask noise suppression, speech modified using a noise vocoder, and speech in reverberation. The new version of HASPI is shown to have a statistically-significant reduction in RMS error compared to the original version for most of the data considered, and to be significantly more accurate for speech in reverberation.

1. Introduction

The Hearing-Aid Speech Perception Index version 1 (HASPI) (Kates and Arehart, 2014) is a metric for predicting monaural speech intelligibility. It is an intrusive metric that requires a reference signal; it compares the time-frequency envelope and temporal fine structure (TFS) of a degraded signal to that of the unprocessed reference. HASPI uses a model of the auditory periphery that can represent impaired as well as normal hearing, and the metric is appropriate for analyzing hearing-aid processing and the effects of hearing loss.

HASPI has successfully been applied to a wide range of problems. It has been used to predict the intelligibility of speech in additive noise (Van Kuyk et al., 2018) and to assess the potential benefit of noise-suppression algorithms for both normal-hearing and hearing-impaired listeners (Kates, 2017). It has also been used to evaluate dynamic-range compression (Rasetshwane et al., 2019), speech-enhancement processing (Hou et al., 2018; Lai and Zheng, 2019), remote microphone systems for hearing aids (Salehi et al., 2018), and to measure the differences between commercial hearing aids (Kates et al., 2018).

The accuracy and range of conditions for which a metric is valid depend on the data used in its derivation, and metrics are often updated to reflect additional datasets (e.g. Rhebergen and Versfeld, 2005; Jensen and Taal, 2016; Steinmetzger et al., 2019). HASPI version 1 was fit to data from four separate experiments. These experiments comprised noise and nonlinear distortion (Kates and Arehart, 2005), frequency compression (Souza et al., 2013; Arehart et al., 2013a), noise suppression (Arehart et al., 2013b; Arehart et al., 2015), and noise-vocoded speech (Anderson, 2010). Reverberant speech was not included in the data used to derive HASPI version 1, and when the metric was applied to recent results for monaural speech intelligibility in reverberation (Muralimanohar, 2018) the accuracy was found to be lower than desired. The objective of this paper is to derive a modified version of HASPI that preserves the accuracy for the original four datasets while improving the accuracy in predicting the intelligibility for speech in reverberation.

A pair of modifications to HASPI is presented in this paper. The first modification is a change to the speech features used to predict intelligibility, and the second is a change to the modelling procedure used to relate the speech characteristics to the listener data. HASPI version 1 uses a 16-ms sliding raised-cosine window filter applied to the time-frequency envelope modulation, so the envelope modulation is lowpass filtered with a cutoff frequency of about 40 Hz. The envelope cross-correlation is combined with a TFS term that cross-correlates the high-intensity portions of the processed and reference signals using 16-ms windowed segments.

In the first modification to produce version 2, the combination of a lowpass envelope filter plus the TFS calculation is replaced by

* Corresponding author.
E-mail address: James.Kates@colorado.edu (J.M. Kates).

https://doi.org/10.1016/j.specom.2020.05.001
Received 4 February 2020; Received in revised form 3 April 2020; Accepted 4 May 2020
Available online 24 May 2020

an envelope modulation filterbank. Envelope modulation filterbanks (Dau et al., 1997) have been successfully used to predict speech intelligibility. The general approach is to pass the signal through an auditory filterbank, extract the envelope in each auditory band, and pass the envelope for each band through a bank of modulation filters (Jørgensen and Dau, 2011; Chabot-Leclerc et al., 2014; Relaño-Iborra et al., 2016; Steinmetzger et al., 2019). The highest modulation rate considered in those papers ranges from a filter center frequency of 64 Hz (Jørgensen and Dau, 2011; Chabot-Leclerc et al., 2014) to 256 Hz (Relaño-Iborra et al., 2016; Steinmetzger et al., 2019). The analysis of Kates and Arehart (2015) indicates that modulation rates below 20 Hz convey the most information for speech intelligibility, but that rates above 64 Hz also make a contribution. The envelope analysis in this paper therefore uses ten bands having center frequencies ranging from 2 to 256 Hz.

HASPI differs from many of the envelope modulation models in that it measures changes in the time-frequency envelope modulation rather than just the envelope changes within each auditory frequency band. Cross-frequency analysis can provide a benefit; for example, Chabot-Leclerc et al. (2014) concluded that incorporating analysis across auditory frequency bands improved the performance of the intelligibility models they studied. Furthermore, Van Kuyk et al. (2018) demonstrated that applying the Karhunen–Loève transform (KLT) (Karhunen, 1947) to the extracted speech features improved prediction performance by decorrelating the model inputs. They point out that the cosine series expansion used in HASPI is similar to the discrete cosine transform (DCT), which is nearly as effective as the KLT in decorrelating the speech analysis features (Ahmed et al., 1974).

The second modification is replacing the parametric model used in version 1 with a neural network (Wasserman, 1989; Beale et al., 2019). The parametric model used in HASPI version 1 assumed *a priori* that a linear combination of the cepstral correlation and TFS features, followed by a sigmoid transformation, was sufficient to accurately model speech intelligibility. The neural network does not make the limiting assumption of a parametric model, and can model an arbitrary nonlinear functional relationship (Beale et al., 2019). A second advantage is that a neural network can model potential interactions between the input variables (Tu, 1996). Interactions between envelope amplitude modulation components occur in the auditory pathway (Joris et al., 2004; Carney, 2018); a neural network allows for potential interactions in forming the intelligibility prediction even if the exact nature of the interactions is not known *a priori*. Overfitting is a potential problem with neural networks (Beale et al., 2019), and this issue is addressed by using bootstrap aggregation (“bagging”) (Breiman, 1996) to generate an ensemble of networks whose outputs are averaged to provide the intelligibility prediction. The ensemble averaging reduces the estimator error variance (Kittler, 1998) and provides improved immunity to overfitting (Krogh and Sollich, 1997; Maclin and Opitz, 1997; Domingos, 2000).

In this paper HASPI version 2 is compared to HASPI version 1. The five listener datasets used for the comparison are first described. The auditory model, envelope analysis, and neural network used for version 2 are then described. The results of the comparison are given, followed by the discussion and conclusions.

2. Intelligibility data

HASPI version 1 was fit to four datasets comprising 1) noise and nonlinear distortion, 2) frequency shifting, 3) noise suppression using an ideal binary mask algorithm, and 4) noise vocoder data. These four datasets are described in Kates and Arehart (2014); further details are provided in the papers cited in each subsection below. In addition to the data used previously, a recent dataset comprising speech in reverberation (Muralimanohar, 2018) has been added. All of these experiments used monaural headphone presentation of the stimuli, and all intelligibility results in this paper are presented as proportion complete sentences correct.

2.1. Noise and distortion data

The noise and distortion dataset is the one analyzed in Kates and Arehart (2005). There were thirteen normal-hearing (NH) adult listeners and nine hearing-impaired (HI) listeners. Intelligibility was scored as sentences correct using the Hearing-in-Noise Test (HINT) materials (Nilsson et al., 1994) spoken by a male talker. Each sentence was combined with additive speech-shaped noise or was processed using symmetric peak clipping or symmetric center clipping. The speech-shaped noise was that provided with the HINT test materials, with signal-to-noise (SNR) values ranging from -5 to 30 dB plus speech in quiet. Thresholds for the peak clipping and center-clipping distortion were derived from the cumulative histogram of the magnitudes of the signal samples for each sentence. The thresholds for peak clipping ranged from infinite clipping to no clipping, and the thresholds for center clipping ranged from 98 percent of the cumulative histogram to no clipping. The sentences for the NH listeners were presented monaurally at 65 dB SPL. For the HI listeners, sentences were amplified using the National Acoustics Laboratories Revised (NAL-R) linear gain rule (Byrne and Dillon, 1986).

2.2. Frequency compression data

The frequency compression dataset is the one analyzed by Souza et al. (2013) and Arehart et al. (2013a). There were fourteen NH adult listeners and twenty-six HI listeners. Intelligibility was scored as sentences correct using the low-context IEEE sentences (Rothauser, 1969) spoken by a female talker. The sentences were combined with multi-talker babble at SNRs ranging from 10 to -10 dB in 5-dB steps or used without any interference, after which the noise-free or noisy speech was processed using frequency compression.

Frequency compression was implemented using sinusoidal modeling (McAulay and Quatieri, 1986) applied to the high-frequency portion of the signal. The speech was first divided into complementary low- and high-frequency bands. The amplitude, phase, and frequency of the ten highest peaks in the high-frequency band were then extracted. The peaks were shifted downwards in frequency, and output sinusoids were synthesized using the preserved amplitude and phase values along with the shifted frequencies (Aguilera Muñoz et al., 1999). The system output comprised the synthesized high-frequency components combined with the unprocessed low-frequency portion of the speech. The parameters for the frequency compression were cutoff frequencies of 1, 1.5, and 2 kHz combined with frequency compression ratios of 1.5:1, 2:1, and 3:1. Also included was a control condition having no frequency compression. The sentences for the NH listeners were presented monaurally at 65 dB SPL. For the HI listeners, sentences were amplified using the NAL-R linear gain rule (Byrne and Dillon, 1986).

2.3. Ideal binary mask noise suppression data

The ideal binary mask (IBM) noise suppression dataset is the one analyzed by Arehart et al. (2013b; 2015). There were seven younger adult NH listeners and thirty older adult HI listeners. Intelligibility was scored as sentences correct using the IEEE sentences (Rothauser, 1969) spoken by a female talker. The sentences were combined with multi-talker babble at SNRs ranging from -18 to 12 dB in steps of 6 dB or used without any interference. The noisy speech was processed through an IBM noise suppression algorithm.

The IBM noise suppression (Kjems et al., 2009; Ng et al., 2013) used a 64-band gammatone auditory filterbank (Patterson et al., 1995). Time frames having a 20-ms duration with a 50 percent overlap were used for the processing. A time-frequency cell is defined as one frame in one frequency band. The local SNR was computed for each cell; if the SNR was 0 or above the cell was kept without attenuation (gain = 0 dB) for the output signal, and if the local SNR was below 0 dB the cell was attenuated using a gain of -10 or -100 dB. In addition, random errors (0, 10, or 30 percent) were introduced into the gain decisions

(Li and Loizou, 2008). The noisy signal was multiplied by the computed gain values to give the processed output in the frequency domain. The processed signal was then filtered through a time-reversed gammatone filterbank and summed across bands to produce the output. The sentences for the NH listeners were presented monaurally at 65 dB SPL. For the HI listeners, sentences were amplified using the NAL-R linear gain rule (Byrne and Dillon, 1986).

2.4. Noise vocoder data

The noise vocoder dataset is the one analyzed by Anderson (2010). There were ten NH adult listeners and ten HI listeners. Intelligibility was scored as sentences correct using the IEEE sentences (Rothauser, 1969) spoken by a male and a female talker. The sentences were combined with multi-talker babble at SNRs of 18 and 12 dB as well being used without any interference. The noise-free or noisy speech was processed through a noise vocoder that encoded a varying number of frequency bands.

The noise vocoder used 32-band linear-phase auditory filterbank. The speech envelope used for the vocoding was extracted via the Hilbert transform and lowpass filtered at 300 Hz. The noise vocoding used Gaussian noise or noise with reduced envelope fluctuations produced by dividing the Gaussian noise by its own envelope in each band. Noise vocoding was applied to the noisy speech beginning with the highest-frequency bands and proceeding downwards in frequency. The vocoding was stepped in groups of two bands from no bands vocoded to the 16 upper frequency bands vocoded, this latter case corresponding to a vocoding cutoff of 1.6 kHz. The speech and vocoded noise bands were then passed through the filterbank a second time to remove out-of-band modulation distortion products and summed across frequency bands to produce the output signal. Finally, the RMS level of the processed signal was matched to that of the input speech. The sentences for the NH listeners were presented monaurally at 65 dB SPL. For the HI listeners, sentences were amplified using the NAL-R linear gain rule (Byrne and Dillon, 1986).

2.5. Reverberation data

The reverberation dataset is the one analyzed by Muralimanohar (2018). The data used are from ten adult NH listeners and nine HI listeners. Intelligibility was scored as sentences correct using the IEEE sentences (Rothauser, 1969) spoken by three male and three female talkers. In the experiments, reverberant speech was processed through several different envelope expansion algorithms to assess their impact on speech intelligibility.

The reverberation used monaural impulse responses recorded in four spaces having T_{60} reverberation times ranging from 627 ms to 3 s. The reverberant speech was passed through a nine-band linear-phase FIR filterbank, and the speech envelope in each band was extracted using the Hilbert transform followed by a 30-Hz linear-phase FIR lowpass filter. Several processing conditions were compared: the clean speech without reverberation, speech with reverberation for the four spaces, noise-vocoded clean and reverberant speech, reverberant speech with its envelope raised to a power of either 1.2 or 2 in all nine bands, reverberant speech with the power in each band chosen to give the minimum mean-squared error (MMSE) match between the envelope of the reverberant speech and the clean speech, reverberant speech with the power in each band chosen to give the MMSE match between the log envelope of the reverberant speech and the clean speech, and reverberant speech with the envelope restored to match that of the clean speech in each band. The speech was then passed through the filterbank a second time to remove out-of-band modulation distortion products and summed across frequency bands to produce the output signal. The sentences for the NH listeners were presented monaurally at 70 dB SPL. For the HI listeners, the 70-dB SPL sentences were amplified using the NAL-R linear gain rule (Byrne and Dillon, 1986).

3. Intelligibility metric

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. 1) 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. 2) The peripheral model outputs comprise the envelopes in each auditory filter band, which are combined across frequency to form time-varying short-time spectra that are then analyzed using a modulation filterbank. 3) The modulation filter outputs are fit to the subject intelligibility data using an ensemble of neural networks. Each of these three processing steps are described below.

3.1. Peripheral model

The model of the auditory periphery is the one used for the original version of HASPI. Descriptions of the model are presented in Kates (2013) and Kates and Arehart (2014) and are summarized here. The overall block diagram for the model is shown in Fig. 1. The model runs at a 24-kHz sampling rate, so sampling rate conversion is provided. Temporal alignment of the reference and degraded signals is implemented in two stages, an initial broadband alignment and a later alignment within each auditory band. The reference and degraded signals are each passed through the peripheral model, described below, after which the envelope modulation features are extracted and compared.

The peripheral model block diagram is presented in Fig. 2. After sampling-rate conversion to 24 kHz, the signal is passed through a middle ear filter (Kates, 1991) which attenuates the low and high frequencies in a manner consistent with equal-loudness contours (Suzuki and Takeshima, 2004). A parallel filterbank comprising fourth-order gammatone filters (Cooke, 1991; Patterson et al., 1995) is used for the auditory frequency analysis. Thirty-two filters cover the frequency range from 80 to 8000 Hz. The filter bandwidths for impaired hearing are increased over those for normal hearing (Moore and Glassberg, 1983) as hearing loss increases (Moore et al., 1999). The filter bandwidths are also increased in response to increasing signal intensity (Baker and Rosen 2002; Baker and Rosen, 2006) for input levels above 50 dB SPL within each frequency band.

The dynamic-range compression mediated by the outer hair cells (Ruggero et al., 1997) is controlled by a separate control filterbank. The bandwidths of the control filters are set to the widest values used in the model (Zhang et al, 2001), and thus correspond to those associated with the maximum hearing loss and/or highest input signal intensity. The envelope of the control filter output is passed through an 800-Hz lowpass filter (Zhang et al, 2001) which provides a small time delay, and the filtered envelope is used to adjust the gain applied to the auditory filter output in response to the signal level in each band. Linear amplification is used for inputs below 30 dB SPL or above 100 dB SPL, and amplitude compression is used for the intermediate input signal levels. Outer hair-cell damage shifts the auditory threshold upwards and reduces the compression ratio. For the maximum hearing loss allowed in the model the amplification thus reverts to a linear system with outputs for signal inputs at 100 dB SPL that match those for the normal ear, thus providing auditory recruitment (Kiessling, 1993).

The auditory signal after compression is converted from the linear amplitude scale to dB re: auditory threshold, with sound levels below the normal or impaired auditory threshold set to a lower limit of 0 dB SPL. Hearing loss due to inner hair-cell damage results in additional attenuation. Modification of the neural firing rates due to inner hair-cell adaptation (Harris and Dallos, 1979) is applied following the dB conversion; a rapid adaptation time constant of 2 ms and a short-term time constant of 60 ms are used. The final stage in the peripheral model is compensation for the time delays associated with the gammatone filters, which corresponds to the timing compensation found in the auditory pathway (Wojtczak et al., 2012).

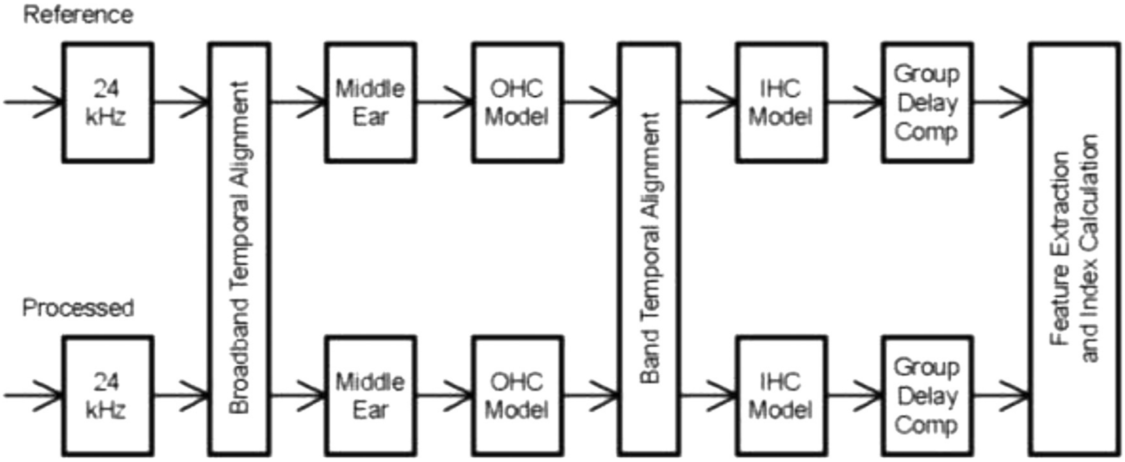


Fig. 1. Block diagram showing the reference and processed signal comparison.

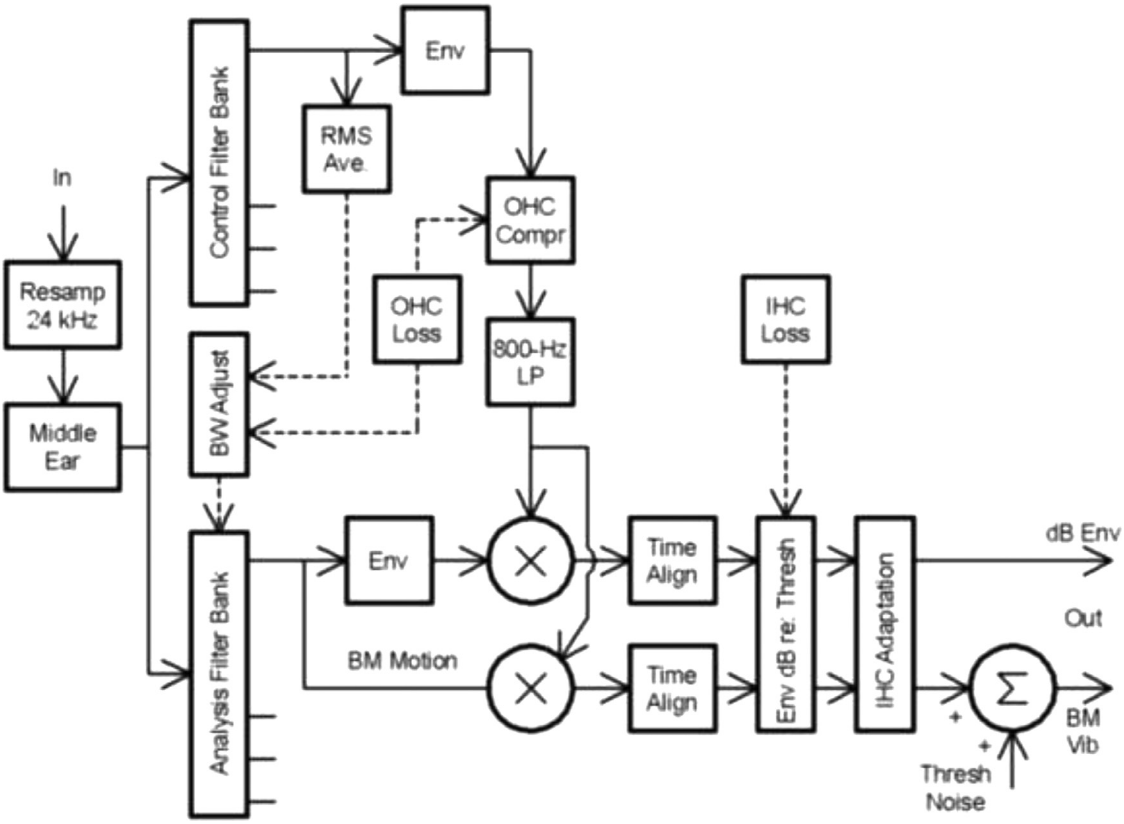


Fig. 2. Block diagram of the auditory model used to extract the signals in each frequency band.

3.2. Envelope modulation analysis

The envelope modulation analysis, shown in the block diagram of Fig. 3, 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 using linear-phase FIR filters having a raised-cosine impulse response corresponding to a cutoff frequency of 320 Hz. The raised-cosine filter response ensures that no negative envelope values are produced. The filtered envelopes are then subsampled at 2560 Hz (8 times the filter cutoff frequency).

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 $\frac{1}{2}$ cycle of a cosine spanning the spectrum from 80 to 8000 Hz and progressing to $2\frac{1}{2}$ cycles spanning the spectrum. The basis functions correspond to the discrete cosine transform (DCT) applied to the short-time dB auditory spectrum, thus giving short-time mel-frequency cepstral coefficients (Mitra et al., 2012). More detail on the cepstral coefficient calculation can be found in Kates and Arehart (2014). The cepstral coefficients also correspond to the principal components of speech determined by Zahorian and Rothenberg (1981); they found that the first six components explained about 92 percent of the short-time spectral variance for male talkers and about 95 percent of the variance for female talkers.

Each of the five cepstral coefficient sequences is passed through a modulation filterbank. The filterbank has ten bands with center fre-

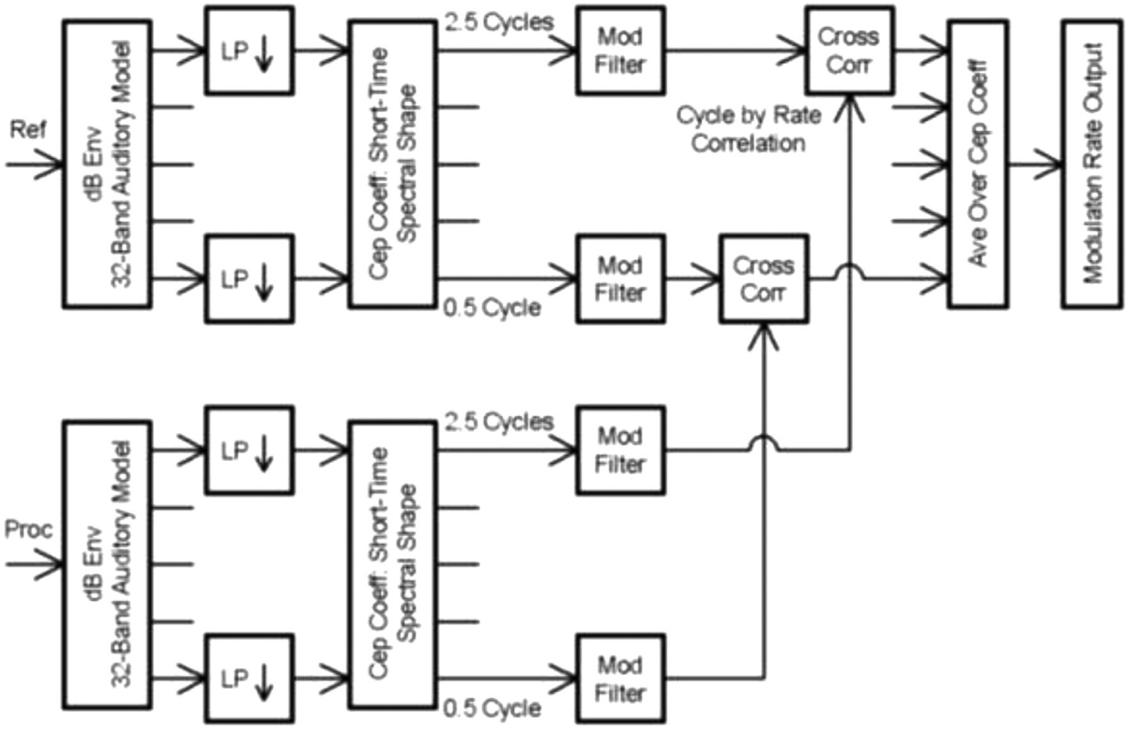


Fig. 3. Block diagram showing the time-frequency envelope modulation analysis. One filter from the modulation filterbank is shown.

Table 1
The ten filters implemented in the modulation filterbank.

Filter No.	Center Frequency, Hz	Lower Edge, Hz	Upper Edge, Hz
1	2	0.0	4.0
2	6	4.0	8.0
3	10	8.0	12.5
4	16	12.5	20.5
5	25	20.5	30.5
6	40	30.5	52.4
7	64	52.4	78.1
8	100	78.1	128.0
9	160	128.0	200.0
10	256	200.0	328.0

quencies ranging from 2 to 256 Hz; the band center frequencies and band edges are indicated in Table 1. 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 (Moritz et al., 2015). The filter responses were adjusted to have a Q of 1.5, which is consistent with the Q values ranging from 1 to 2 that have been fit to amplitude modulation discrimination data (Dau et al., 1997; Ewert et al. 2000; Ewert et al., 2002).

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 for the five DCT basis functions are similar to each other (Kates and Arehart, 2015), and the cross-correlation values were averaged across the basis functions to produce an output vector comprising the averaged covariances for the ten modulation filters.

3.3. Neural network ensemble

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 (Beale et al. 2019) and for its ability to model of potential interactions between the input variables (Tu, 1996). 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 (Rumelhart et al., 1986; Wasserman, 1989; Werbos, 1990). The networks were implemented using custom MATLAB code, and the functions needed to run HASPI version 2 are provided as part of the code distribution.

The ten neural networks were each initialized to a different independent set of random weights, and the models were fit to the concatenation of the individual subject scores from the five experiments. The NH and HI subject groups and each of the five datasets were assigned comparable importance in fitting the networks by replicating the subject data an integral number of times to provide approximately 5000 data points for the NH and 5000 points for the HI listener groups for each experiment. The combined dataset comprised 72166 sample vectors including the data replication, and fifty iterations of the dataset with replications were used to train each neural network in the ensemble.

The ensemble averaging approach was chosen to reduce the possibility of overfitting the data. Overfitting occurs when the network has enough free parameters that it can learn the details of a specific dataset rather than provide a general solution (Beale et al., 2019). To reduce the possibility of overfitting, bootstrap aggregation (“bagging”) was used (Breiman, 1996) in which the outputs from several parallel networks are averaged. Each neural network was trained using a subset comprising (1-1/e) = 0.632 of the data, selected with replacement (Efron and Gong, 1983; Breiman, 1996). The bagging approach gives reduced estimator error variance (Kittler, 1998) and provides improved immunity to overfitting (Krogh and Sollich, 1997; Maclin and Opitz, 1997; Domingos, 2000).

The network training for ensemble averaging differs from what is generally recommended when only a single neural network is used to model the data. A procedure such as *k*-fold cross validation is often used to assess the accuracy of a single network, and a separate validation dataset used to determine the optimum stopping point for the network weight adaptation (Beale et al, 2019). For an ensemble, however, the accuracy of the average across the networks is actually enhanced when there is disagreement among the networks as can occur when individual models are overfit (Naftaly et al., 1997; Krogh and Sollich, 1997; Cunningham et al., 2000), so using the conventional optimal stopping point for the adaptation can actually be counter-productive. An average of ten neural networks is sufficient to provide the main benefits of bagging in reducing overfitting (Hansen and Salamon, 1990; Breiman, 1996; Opitz and Maclin, 1999), so an ensemble of ten networks was used for the revised intelligibility model.

4. Results

4.1. Intelligibility plots

Scatter plots for the intelligibility predictions averaged over subjects are presented in Fig. 4. The plots for HASPI version 2 are in the left column, and the plots for version 1 are in the right column. Each point in the plots represents one processing condition; the *x*-axis coordinate is the average of the model predictions over the subjects in the group and the y-axis coordinate is the average of the subject intelligibility scores for the members of the group. The NH listener group is identified by the open circles, and the HI listener group is identified by the filled squares. The diagonal line shows perfect performance; a point above the line indicates that the predicted intelligibility is lower than the subject scores, and a point below the line indicates that the predicted intelligibility is higher than the subject scores. A low RMS error will result in the distribution being close to the line. A high degree of correlation between the predicted and observed values will result in the distribution of points along a line, but the line will not necessarily be congruent with the 45-degree line drawn in the plot.

The scatter plots for the noise and distortion dataset are presented in the first row of Fig. 4. These plots are similar for the two versions of HASPI. For both plots, a large number of points are concentrated in the vicinity of (1,1), which represents perfect intelligibility. These points, when combined with the points near (0,0), exert a strong influence on the correlation coefficients between the models and the subject data and ensure similar performance. In addition, both version 1 and version 2 have similar numbers of points above and below the diagonal, indicated little apparent bias in the predictions.

The scatter plots for the frequency compression dataset are presented in the second row of Fig. 4. For both version 1 and version 2, the majority of points for the NH listeners are plotted above the diagonal line while those for the HI listeners are plotted below the line, thus showing a slight bias towards underestimating the intelligibility for the NH listeners and overestimating the intelligibility for the HI listeners. There are also outliers along the lower right edge of the version 1 plot that are absent from the version 2 plot. These points correspond to speech in quiet where the cutoff frequency for the frequency compression has been set to 1 kHz. These outliers in the version 1 predictions, where the model predicts much higher intelligibility than observed for the subjects, suggest that the envelope modulation differences detected by the lowpass filter used in version 1 cannot respond accurately to the envelope changes introduced by the frequency compression algorithm, but that the modulation filterbank used in version 2 can detect these changes.

The scatter plots for the ideal binary mask noise suppression data are presented in the third row of Fig. 4. The points for the NH listeners all lie below the diagonal for version 1, but are closer to the diagonal in version 2 with some NH points now above the diagonal. However, all of the HI points lie below the diagonal for version 2 even though some points lie above the diagonal for version 1. Both version 1 and version 2 have a large number of points near the origin and near (1,1), so these points will tend to cause high correlation coefficients between the model and observed scores independent of the distributions of the points near the center of the plots.

The scatter plots for the reverberation dataset are presented in the bottom row of Fig. 4. All of the points for version 1 lie below the diagonal, indicating a large bias in the version 1 predictions. HASPI version 1 consistently overestimates the intelligibility for speech in reverberation for both NH and HI listeners. There are also a large number of HI points near the right edge of the version 1 plot where version 1 indicates intelligibility of nearly 100 percent sentences correct, but where the listener performance is as low as 30 percent correct. The bias in the predictions and the number of outliers are both substantially reduced for version 2.

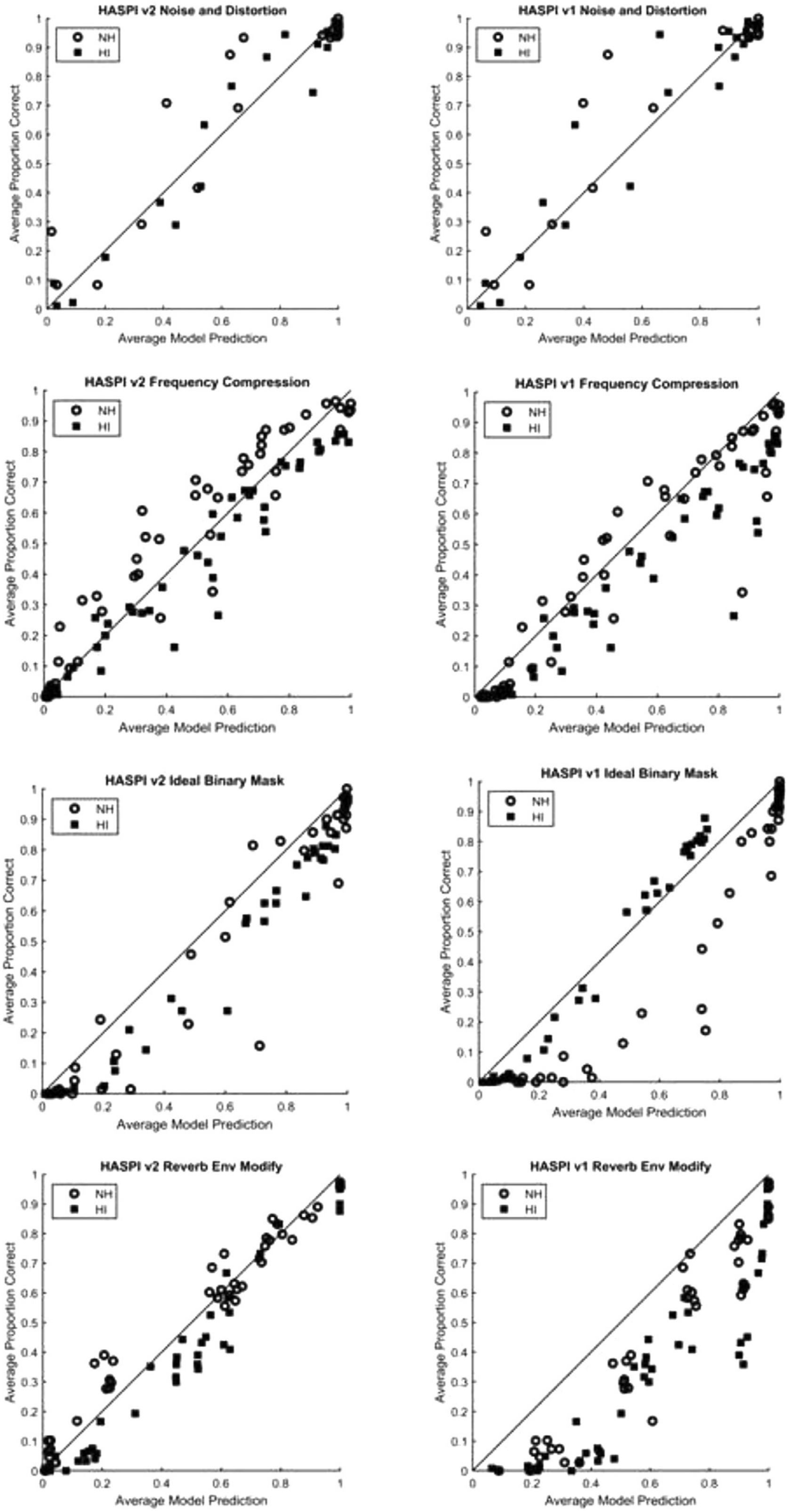
The scatter plots for the noise vocoder dataset are presented in Fig. 5. This dataset was included to provide situations where the speech TFS was corrupted but where the modification had only a small impact on intelligibility, thus reducing any undue dependence of the metric on changes in the signal TFS. Note that since all the processing conditions yield very high observed and predicted intelligibility, the x- and y-axes have been expanded to cover the range 0.75 to 1. For the NH listeners, all the combinations of SNR and TFS removal gave nearly perfect intelligibility, and both versions of HASPI give predicted values very close to 1 as well. Since there are no consistent differences due to processing condition, the remaining scatter in the data is most likely due to individual subject variability. For the HI listeners, the HASPI predictions are concentrated in vertical bands based on SNR with minimal effect of the TFS removal. The spread in the HI listener responses is somewhat larger than those for the NH listeners, possibly due to the reduction in the ceiling effect on the intelligibility scores.

4.2. Differences between versions 1 and 2

The performance for HASPI version 1 and for version 2 was computed using a non-parametric bootstrap procedure (Efron, 1983; Efron and Gong, 1983; Zio, 2006) based on the MATLAB functions provided by Rousselet (2017). All calculations used 10000 bootstrap replications (with replacement) of the data to estimate the probability density functions from which the means, standard deviations, and confidence intervals were calculated. Comparisons of the models comprise the RMS error between the model outputs and the listener responses, the Pearson correlation coefficient, Spearman rank-order correlation coefficient, and Kendall's tau coefficient for pair-wise comparisons. The RMS error indicates how closely the model outputs agree with the subject scores; it is sensitive to both the bias and variability in the model outputs and is sensitive to outliers. The Pearson correlation coefficient measures the degree of linearity between the predicted and observed scores and is also sensitive to outliers. The Spearman correlation assesses the degree to which the rank ordering of the processing conditions matches between the predicted and observed scores, and Kendall's tau indicates the accuracy in identifying the better of each pair of processing comparisons. Both the Spearman and Kendall correlation coefficients are non-parametric and have reduced sensitivity to outliers since they depend only on the rankings and not on the magnitude of the differences.

The results for HASPI version 1 are presented in Table 2 for the five datasets. The entries in this table differ slightly from those in Kates and Arehart (2014) due to the use of bootstrapping to compute the values in the present paper. In general, the correlations and RMS error values are consistent in that the highest correlations are associated with the smallest RMS errors. There are, however, some data for which the RMS error is high even though the Pearson correlation is also high, examples being the IBM noise suppression for NH listeners and the reverberation experiment, which reflects the bias visible in the scatterplots of Fig. 4.

The noise vocoder results show RMS errors that are substantially lower than for the other datasets. The low RMS errors are consistent


HASPI v2 Noise and Distortion

HASPI v1 Noise and Distortion

HASPI v2 Frequency Compression

HASPI v1 Frequency Compression

HASPI v2 Ideal Binary Mask

HASPI v1 Ideal Binary Mask

HASPI v2 Reverb Env Modify

HASPI v1 Reverb Env Modify

Fig. 4. Scatterplots comparing HASPI v2 (left column) to HASPI v1 (right column). Each point represents the model prediction compared to the listener intelligibility scores for a processing condition averaged over all listeners for that condition. Data for normal-hearing (NH) listeners is plotted using the open circles and that for hearing-impaired (HI) listeners is plotted using filled squares. The different experiments are identified in the figure titles.

with the scatter plots of Fig. 5, since both the listener scores and the HASPI predictions are close to 1 for all processing conditions. The scatter plots of Fig. 5 also show that the intelligibility scores plotted against the HASPI predictions for the NH and HI subject groups form two non-overlapping clusters. The correlation coefficients are low for each group because the intelligibility scores are all similar to each other and HASPI cannot model the remaining inter-subject variability. However, combining the NH and HI data creates a dumbbell-shaped distribution comprising the NH and HI clusters, which is sufficient to give much higher correlations reflecting the separation of the cluster centers.

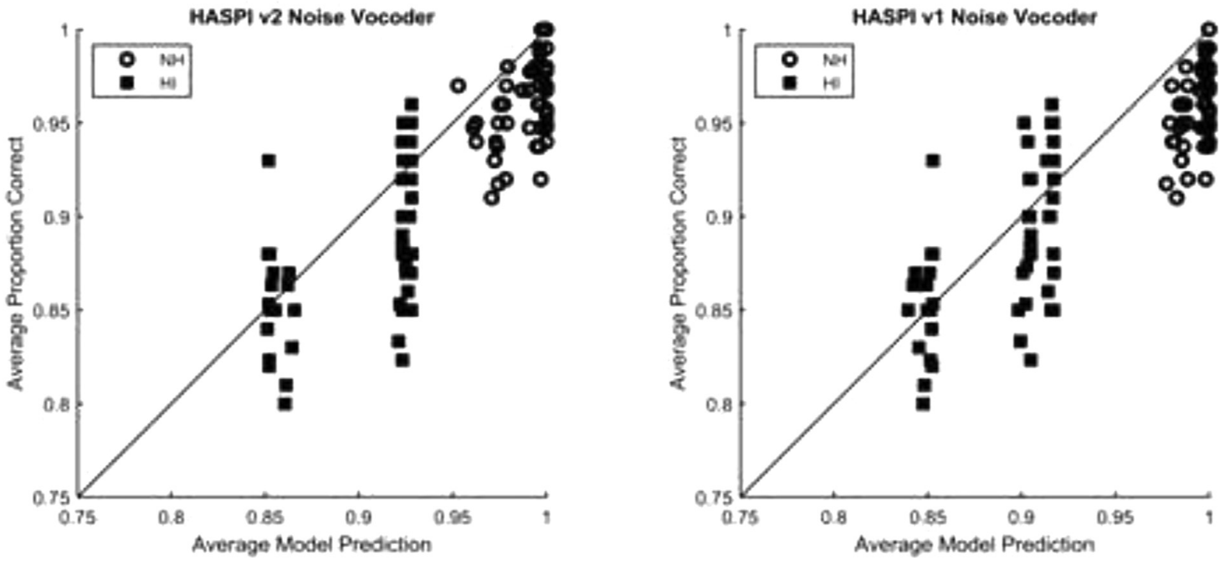


Fig. 5. Scatterplots comparing HASPI v2 (left plot) to HASPI v1 (right plot) for the noise vocoder dataset. Note that the x- and y-axes span 0.75 to 1.

Table 2
RMS error and correlations between the model predictions and the listener responses for HASPI version 1. The results are averaged over the listeners in each hearing group. The RMS error and correlation values are the mean of 10000 bootstrap replications of the model output.

Experiment	Subj Group	RMS Error	Pearson	Spearman	Kendall
Noise and Distortion	NH	.1134	.9344	.8539	.7308
	HI	.0918	.9600	.8919	.7794
	NH+HI	.1050	.9486	.8726	.7255
Freq Compression	NH	.1048	.9636	.9615	.8660
	HI	.1469	.9662	.9575	.8495
	NH+HI	.1289	.9576	.9552	.8322
IBM Noise Suppress	NH	.2039	.9529	.9530	.8466
	HI	.0633	.9935	.9820	.9280
	NH+HI	.1511	.9417	.9417	.8269
Noise Vocoder	NH	.0379	.4933	.4237	.3125
	HI	.0344	.5716	.5777	.4264
	NH+HI	.0363	.8445	.8269	.6425
Reverb Env Modify	NH	.1872	.9605	.9307	.7939
	HI	.2797	.9186	.9445	.8231
	NH+HI	.2382	.9346	.9410	.8053

The results for HASPI version 2 are presented in Table 3 for the five datasets. The overall results for version 2 are similar to those for version 1, but there are some differences. In particular, the RMS error for the IBM noise suppression for NH listeners and the reverberation experiment is greatly reduced and the Pearson correlations are higher.

The comparison between the two versions of HASPI is presented in Table 4. The performance metric for version 1 is subtracted from that for version 2. A *negative* RMS error difference thus indicates a larger error for version 1 than for version 2, so version 2 is more accurate. A *positive* difference in a correlation coefficient indicates a higher degree of correlation for version 2 than for version 1, indicating that version 2 is better. Differences that are significant for rejecting the null hypothesis of identical performance between the two versions are identified in bold italics, with one asterisk for statistical significance at the 5 percent level and two asterisks for significance at the 1 percent level. Significance was computed by generating the distribution of the model differences for the 10000 bootstrap replications. The mean difference in the bootstraps is presented in Table 4, and the tails of the distributions were used to estimate the appropriate confidence intervals for establishing statistical significance.

Most of the differences in RMS error are negative, indicating that version 2 has reduced error compared to version 1. The differences are not statistically significant at the five-percent level for the noise and distortion dataset, but there are statistically-significant differences in RMS error for the other four datasets. Version 2 has significantly reduced RMS error for the HI and combined subject groups for frequency compression, significantly reduced RMS error for the NH and combined groups for noise suppression, significantly reduced RMS error for the NH group for the noise vocoder experiment, and significantly reduced RMS error for all groups for the reverberation dataset. The RMS error was significantly increased for the noise suppression HI group and noise vocoder HI group. There were significant increases in the Pearson correlation coefficient for the noise suppression dataset NH and combined subject groups and the reverberation dataset for all subject groups. The Pearson correlation coefficient was significantly reduced for version 2 for the noise vocoder HI group. There were also statistically-significant increases for version 2 in the Spearman correlation for the frequency compression dataset HI group and in the Kendall correlation coefficient for the frequency compression HI group and reverberation combined group. The improvements for the reverberation dataset in both RMS error and Pearson correlation coefficient indicate both a reduction in the bias and number of outliers for version 2 compared to version 1.

The effect sizes (Cohen's *d*) (Sullivan and Feinn, 2012) for the significant differences are presented in Table 5. The differences in the means

Table 3
RMS error and correlations between the model predictions and the listener responses for HASPI version 2 fit to the combined normal-hearing (NH) and hearing-impaired (HI) data. The results are averaged over the listeners in each hearing group. The RMS error and correlation values are the mean of 10,000 bootstrap replications of the model output.

Experiment	Subj Group	RMS Error	Pearson	Spearman	Kendall
Noise and Distortion	NH	.1125	.9390	.8284	.7089
	HI	.0764	.9737	.9027	.7791
	NH+HI	.0967	.9562	.8870	.7381
Freq Compression	NH	.0985	.9691	.9723	.8855
	HI	.0806	.9809	.9698	.8770
	NH+HI	.0906	.9657	.9638	.8470
IBM Noise Suppress	NH	.1219	.9666	.9333	.8084
	HI	.1160	.9863	.9730	.9038
	NH+HI	.1199	.9739	.9585	.8515
Noise Vocoder	NH	.0338	.4482	.4428	.3251
	HI	.0413	.5255	.5131	.3706
	NH+HI	.0378	.8296	.8224	.6318
Reverb Env Modify	NH	.0661	.9851	.9609	.8569
	HI	.0932	.9763	.9640	.8697
	NH+HI	.0811	.9720	.9622	.8542

Table 4
Differences between the RMS error and correlations computed as the HASPI version 2 value minus the HASPI version 1 value. A negative difference in the RMS error indicates that version 2 has reduced error compared to version 1. A positive difference in the correlations indicates that version 2 is more highly correlated with the subject data than version 1. The RMS error and correlation differences are the average of 10,000 bootstrap replications of the differences in the model outputs, and significance was computed from the bootstrapped confidence intervals. Differences at the 5 percent level are indicated by bold italics plus one asterisk, while differences at the 1 percent level are indicated by bold italics plus two asterisks.

Experiment	Subj Group	RMS Error	Pearson	Spearman	Kendall
Noise and Distortion	NH	-0.0009	0.0046	-0.0255	-0.0219
	HI	-0.0154	0.0137	0.0108	-0.0003
	NH+HI	-0.0083	0.0075	0.0145	0.0126
Freq Compression	NH	-0.0064	0.0055	0.0108	0.0195
	HI	***-0.0662*****	0.0147	***0.0123****	***0.0275****
	NH+HI	***-0.0382*****	0.0081	0.0086	0.0147
IBM Noise Suppress	NH	***-0.0820*****	***0.0138****	-0.0196	-0.0382
	HI	***0.0526*****	-0.0071	-0.0090	-0.0242
	NH+HI	***-0.0312****	***0.0322*****	0.0168	0.0246
Noise Vocoder	NH	***-0.0040*****	-0.0451	0.0190	0.0127
	HI	***0.0069*****	***-0.0461****	-0.0646	-0.0558
	NH+HI	0.0016	-0.0149	-0.0044	-0.0107
Reverb Env Modify	NH	***-0.1211*****	***0.0246*****	0.0302	0.0630
	HI	***-0.1864*****	***0.0606*****	0.0195	0.0467
	NH+HI	***-0.1571*****	***0.0373*****	0.0212	***0.0489****

Table 5
Effect sizes (Cohen's *d*) for the statistically significant model differences identified in Table 4. The effect sizes were computed from the bootstrapped means and standard deviations.

Experiment	Subj Group	RMS Error	Pearson	Spearman	Kendall
Noise and Distortion	NH	—	—	—	—
	HI	—	—	—	—
	NH+HI	—	—	—	—
Freq Compression	NH	—	—	—	—
	HI	3.8079	—	0.9861	1.0451
	NH+HI	3.1041	—	—	—
IBM Noise Suppress	NH	3.0292	—	—	—
	HI	5.6765	1.7187	—	—
	NH+HI	1.8279	2.5396	—	—
Noise Vocoder	NH	1.4931	—	—	—
	HI	1.9656	0.4885	—	—
	NH+HI	—	—	—	—
Reverb Env Modify	NH	9.8800	2.8392	—	—
	HI	12.6737	4.3429	—	—
	NH+HI	14.7859	3.6029	—	2.0688

and the standard deviations were computed from the bootstrap distributions used for Table 4, and Cohen's *d* was calculated as the difference in the means divided by the pooled standard deviation. An effect size of 0.2 is considered to be small, 0.5 medium, 0.8 large, and 1.3 very large. Based on this classification, the effect size for the noise vocoder HI group is small to medium, while all other effect sizes qualify as large or greater. The greatest effect sizes are for the RMS error and Pearson correlation coefficients for the reverberation dataset, where all effect sizes exceed the very large category definition by a substantial margin.

5. Discussion

5.1. Modulation Filter Weights

It is difficult to interpret the neural network weights, in particular the relative importance of the different modulation-rate filters in forming the model of the listener intelligibility scores. To gain some insight into which modulation rates are most important, two parametric models were also fit to the subject data using a minimum mean-squared error criterion. While the accuracy of the parametric models fell short of that of the neural network, the relative weights assigned to the ten modulation-rate filter outputs provide some indication of their relative importance.

The first model is similar to the parametric fit approach used in HASPI version 1. The cepstral sequences for basis functions 2 through 5 were each passed thorough the modulation filterbank as shown in Fig. 1 and averaged across the cepstral basis functions as was done for the neural network model. A sigmoid transformation was then applied to a weighted sum of the ten modulation filter outputs. The predicted intelligibility I is given by:

$$p = b + \sum_{j=1}^{10} w_j c_j$$
$$I = \frac{1}{1+e^{-p}} \quad (1)$$

where c_j is the averaged cepstral correlations computed for modulation filter j, w_j are the weights, and b is a bias term. A second transformation was also investigated, this being a saturating power law transformation that was found to provide better accuracy than the sigmoid function. For the saturating power law, the predicted intelligibility I is given by:

$$q = \sum_{j=1}^{10} w_j c_j$$
$$I = \frac{q^a}{1+q^a} \quad (2)$$

where an exponent of $a = 5.58$ was empirically found to give the best match of the model to the subject data.

The weights w_j for each of the two parametric models are presented in Table 6. Both model fits show similar patterns in that the largest weights are assigned to modulation rates for 10 Hz and below and for 64 Hz and above, with reduced importance for modulation rates between 16 and 40 Hz. The greater dependency on the low and high modulation rates is consistent with the general dependencies of the original version of HASPI, in which intelligibility was found to be a function of lowpass-filtered envelope modulation below about 40 Hz combined with the TFS term. The dependency also agrees with the results of Steinmetzger et al. (2019), who found that modulation filters tuned to intermediate modulation rates made the smallest contribution to their intelligibility model. These intermediate modulation rates correspond to the sensation of auditory roughness, which is highest for modulation rates between 50 and 100 Hz for both NH (Fastl, 1974) and HI listeners (Tufts and Molis, 2007). It is hypothesized that little relevant speech information is contained in the roughness modulation region (Joris et al, 2004; Arnal et al., 2015); prosodic and phonemic information is concentrated at rates below 50 Hz (Rosen, 1992; Mesgarani et al., 2008), while pitch periodicity is associated with rates above 100 Hz (Goy et al., 2013).

Table 6
Band weights for parametric function fits of the modulation filterbank outputs to the subject data.

Filter No.	Center Frequency, Hz	Sigmoid Fit	Power Law Fit
1	2	1.361	0.277
2	6	1.521	1.028
3	10	1.164	0.404
4	16	0.492	0.093
5	25	0.436	0
6	40	0.690	0
7	64	1.142	0
8	100	0.816	1.279
9	160	1.576	0
10	256	2.269	0

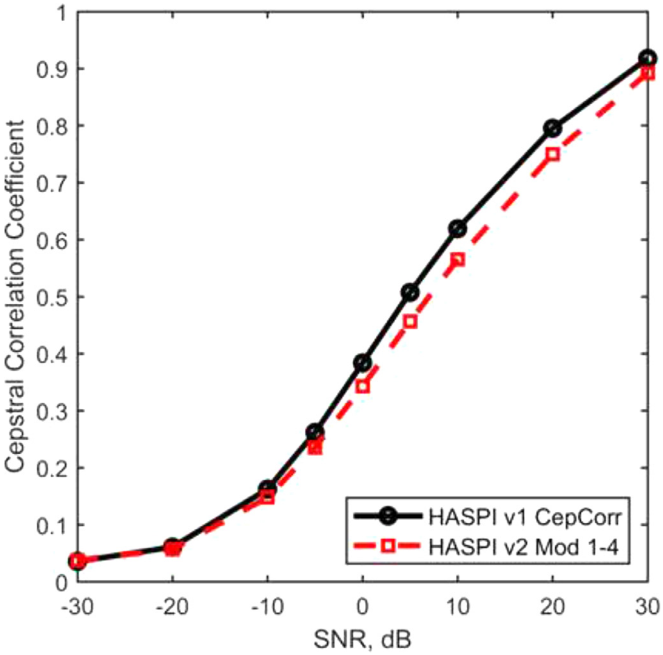


Fig. 6. HASPI version 1 cepstral correlation (solid black line) and HASPI version 2 modulation filterbank outputs averaged over cepstrum basis functions 2–6 and then averaged over the four lowest-rate modulation filters (red dashed line) as a function of SNR for speech in babble.

5.2. Comparing speech measurements

HASPI version 1 measures the cepstral correlation by passing the sequence of cepstral coefficients through a lowpass filter and averaging over coefficients 2 through 6, and measures the TFS by averaging the segmental cross-covariance computed for the high-level speech segments. HASPI version 2 uses the cepstral coefficient sequences passed through a modulation filterbank followed by averaging over coefficients 2 through 6. These two sets of measurements are closely related.

The raw measurements for HASPI versions 1 and 2 were computed for twenty concatenated sentences taken from the IEEE sentence lists (Rothauser, 1969). Two sentences were randomly selected from each of ten talkers; five of the talkers were female and five male. The sentences were combined with the six-talker babble from the Connected Speech Test (CST) (Cox et al., 1987) to give noisy speech at SNRs ranging from -30 to +30 dB. The RMS levels of all stimuli were set to 65 dB SPL, and normal hearing was assumed.

In Fig. 6, the version 1 lowpass filtered cepstral correlation is plotted as a function of SNR using the solid black line. The version 2 modulation

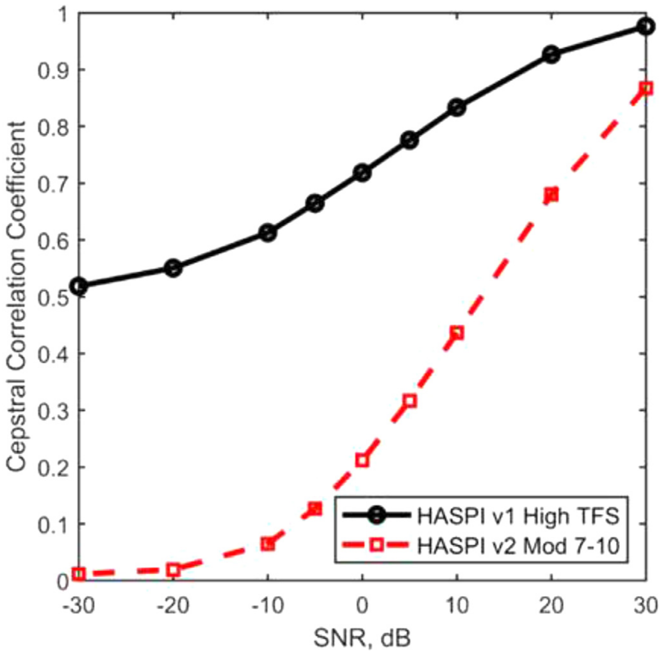


Fig. 7. HASPI version 1 high-intensity segmental TFS correlation (solid black line) and HASPI version 2 modulation filterbank outputs averaged over cepstrum basis functions 2-6 and then averaged over the four highest-rate modulation filters (red dashed line) as a function of SNR for speech in babble.

filterbank outputs averaged over the four lowest-rate bands are plotted using the dashed red line. The four modulation rate filters have center frequencies of 2, 6, 10, and 16 Hz, and span modulation rates from 0 to 20.5 Hz. The two curves have very similar shapes, and the Pearson correlation coefficient between the two curves is 0.999. In Fig. 7, the version 1 high-level TFS term is plotted as a function of SNR using the solid black line. The version 2 modulation filterbank outputs averaged over the four highest-rate bands are plotted using the dashed red line. The four modulation rate filters have center frequencies of 64, 100, 160, and 256 Hz, and span modulation rates from 52.4 to 328 Hz. The high-level TFS curve spans about half the range spanned by the averaged modulation filter outputs, but the curves are still closely related to each other since they have a Pearson correlation coefficient of 0.972.

The high degree of correlation between the two sets of curves indicates that they respond to similar changes in the processed signals. The lower four modulation rate filters used in version 2 are included in the lowpass filter used for version 1, so there is a large degree of overlap in the envelope modulation measurements. The version 1 TFS term is based on the cross-correlation of the high-intensity degraded and reference signals in 16-ms segments, so envelope modulation at rates corresponding to the upper four modulation rate filters used in version 2 is subsumed into the version 1 TFS calculation. Both the TFS and higher-rate modulation filter outputs respond to changes in the speech periodicity, which Steinmetzger et al. (2019) argue aids in predicting the effects of periodic maskers. The correspondence between the TFS and envelope modulation is also consistent with the ability of the auditory system to extract envelope information from the TFS (Gitza, 2001; Zeng et al., 2004; Apoux et al., 2011).

6. Conclusions

The goal of this paper was to develop a new version of HASPI that was more accurate for speech intelligibility in reverberation but which did not sacrifice accuracy for the signal modifications used in deriving the original version. The revised version of HASPI uses a filterbank analysis of the envelope time-frequency modulation combined with an ensemble of neural networks to match the speech measurements to the listener data. The revised metric significantly reduces the RMS error in predicting intelligibility compared to the original version for a majority of speech processing conditions considered. It significantly improves both the Pearson correlation and the RMS error for speech in reverberation, thus achieving the processing objectives.

In the original HASPI paper (Kates and Arehart, 2014), the authors showed that a speech intelligibility metric that combined envelope and TFS measurements was more accurate than on based on TFS or envelope modulation alone. The results in this paper demonstrate that a metric that replaces the TFS measurement with time-frequency envelope modulations at high modulation rates is also effective in providing accurate intelligibility predictions.

The original version of HASPI used a parametric model to compute the intelligibility prediction. The model comprised a weighted combination of the cepstral correlation and high-intensity TFS correlation, followed by a sigmoid transformation. The revised version replaces the parametric model with an ensemble of neural networks. This approach is more general, allowing for interactions between the modulation rates and an arbitrary mapping of the inputs to the subject intelligibility scores, and the bootstrap aggregation helps prevent the loss of generalization that can sometimes occur in statistical data modelling.

The MATLAB code for HASPI version 2 (as well as version 1) is available upon request from the first author at James.Kates@colorado.edu

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors were supported by a research grant to the University of Colorado from GN ReSound.

References

Aguilera Muñoz, C.M., Nelson, P.B., Rutledge, J.C., Gago, A., 1999. Frequency lowering processing for listeners with significant hearing loss. Electron. Circuits Syst: Proc. 2, 741–744 ICECS 1999Pafos, Cypress, Sept 5-8, 1999.

Ahmed, N., Natarajan, T., Rao, K.R., 1974. The discrete cosine transform. IEEE Trans. Comput. C-23, 90–93.

Anderson, M.C. 2010. The role of temporal fine structure in sound quality perception. Speech, Language, and Hearing Sciences Graduate Theses & Dissertations 3. Downloaded from https://scholar.colorado.edu/concern/graduate_thesis_or_dissertations/j9602061v. (Last viewed 28 January 2020).

Apoux, F., Millman, R.E., Viemeister, N.F., Brown, C.A., Bacon, S.P., 2011. On the mechanisms involved in the recovery of envelope information from temporal fine structure. J. Acoust. Soc. Am. 130, 273–282.

Arehart, K.H., Souza, P., Baca, R., Kates, J.M., 2013a. Working memory, age, and hearing loss: susceptibility to hearing aid distortion. Ear Hear. 34, 251–260.

Arehart, K.H., Souza, P.E., Lunner, T., Pedersen, M.S., Kates, J.M., 2013b. Relationship between distortion and working memory for digital noise-reduction processing in hearing aids. In: Proc. Mtgs. Acoust. (POMA) 19, 050084: Acoust. Soc. Am. 165th Meeting. Montreal June 2–7, 2013.

Arehart, K.H., Souza, P.E., Kates, J.M., Lunner, T., Pedersen, M.S., 2015. Relationship among signal fidelity, hearing loss, and working memory for digital noise suppression. Ear Hear. 36, 505–516.

Arnal, L.H., Flinker, A., Kleinschmidt, A., Giraud, A-L., Poeppel, D., 2015. Human screams occupy a privilidged niche in the communication soundscape. Curr. Biol. 25, 2051–2056.

Baker, R.J., Rosen, S, 2002. Auditory filter nonlinearity in mild/moderate hearing impairment. J. Acoust. Soc. Am. 111, 1330–1339.

Baker, R.J., Rosen, S, 2006. Auditory filter nonlinearity across frequency using simultaneous notch-noise masking. J. Acoust. Soc. Am. 119, 454–462.

Beale, M. H., Hagan, M. T., and Demuth, H. B., 2019. Deep Learning Toolbox: User's Guide, R2019b. Downloaded from https://www.mathworks.com/help/pdf_doc/deeplearning/nnet_ug.pdf (Last viewed 6 November 2019).

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Byrne, D., 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.

Carney, L.H., 2018. Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss. J. Assn. Res. Otolaryng. 19, 331–352.

Chabot-Leclerc, A., Jørgensen, S., Dau, T., 2014. The role of auditory spectro-temporal modulation filtering and the decision metric for speech intelligibility prediction. J. Acoust. Soc. Am. 135, 3502–3512.

Cooke, M., 1991. Modeling auditory processing and organization Ph.D. Thesis. U. Sheffield.

Cox, R., Alexander, G., Gilmore, C., 1987. Development of the Connected Speech Test (CST). Ear Hear. 8, 119S–125S Suppl.

Cunningham, P., Carney, J., Jacob, S., 2000. Stability problems with artificial neural networks and the ensemble solution. Artif. Intell. Med. 20, 217–225.

Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modelling auditory processing of amplitude modulation. I: Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102, 2892–2905.

Domingos, P., 2000. Bayesian averaging of classifiers and the overfitting problem. In: Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, pp. 223–230.

Efron, B., 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. J. Am. Stat. Assn. 78, 316–331.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am. Stat. 37, 36–48.

Ewert, S.D., Dau, T., 2000. Characterizing frequency selectivity for envelope fluctuations. J. Acoust. Soc. Am. 108, 1181–1196.

Ewert, S.D., Verhey, J.L., Dau, T., 2002. Spectro-temporal processing in the envelope-frequency domain. J. Acoust. Soc. Am. 112, 2921–2931.

Fastl, H., 1974. The hearing sensation roughness and neuronal responses to AM-tones. Hear. Res. 46, 293–296.

Gitza, O., 2001. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J. Acoust. Soc. Am. 110, 1628–1640.

Goy, H., Fernandes, D.N., Pichora-Fuller, M.K., van Lieshout, P., 2013. Normative voice data for younger and older adults. J. Voice 27, 545–555.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. 12, 993–1001.

Harris, D.M., Dallos, P., 1979. Forward masking of auditory nerve fiber responses. J. Neurophys. 42, 1083–1107.

Hou, J-C., Wang, S-S., Lai, Y-H., Tsao, Y., Chang, H-W., Wang, H-M., 2018. Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Trans. Emerg. Top. Comput. Intel 2, 117–128.

Jensen, J., Taal, C.H., 2016. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. IEEE/ACM Trans. Audio Speech Lang. Proc. 24, 2009–2022.

Jørgensen, S., Dau, T., 2011. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. 130, 1475–1487.

Joris, P.X., Schreiner, C.E., Rees, A., 2004. Neural processing of amplitude-modulated sounds. Physiol. Rev. 84, 541–577.

Karhunen, K., 1947. Über lineare Methoden in der Wahrscheinlichkeitsrechnung (On Linear Methods in Probability and Statistics). Helsinki, Finland: Universitat Helsinki, 1947.

Kates, J.M., 1991. A time domain digital cochlear model. IEEE Trans. Sig. Proc. 39, 2573–2592.

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

Kates, J.M., 2017. Modeling the effects of single-microphone noise suppression. Speech Comm. 90, 15–25.

Kates, J.M., Arehart, K.H., 2005. Coherence and the speech intelligibility index. J. Acoust. Soc. Am. 117, 2224–2237.

Kates, J.M., Arehart, K.H., 2014. The Hearing-Aid Speech Perception Index (HASPI). Speech Comm. 65, 75–93.

Kates, J.M., Arehart, K.H., 2015. 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., Arehart, K.H., Anderson, M.C., Muralimanohar, R.Kumar, Harvey Jr, L.O., 2018. Using objective metrics to measure hearing aid performance. Ear Hear. 39, 1165–1175.

Kiessling, J., 1993. Current approaches to hearing aid evaluation. J. Speech-Lang. Path. Audiol. Monogr. Suppl. 1, 39–49.

Kittler, J., 1998. Combining classifiers: A theoretical framework. Pattern Anal. Appl. 1, 18–27.

Kjems, U., Boldt, J.B., Pedersen, M.S., Wang, D., 2009. Role of mask pattern in intelligibility of ideal binary-masked noisy speech. J. Acoust. Soc. Am. 126, 1415–1426.

Krogh, A., Sollich, P., 1997. Statistical mechanics of ensemble learning. Phys. Rev. E 55, 811–825.

Lai, Y-H., Zheng, W-Z., 2019. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing-device users. Biomed. Sig. Proc. Control 48, 35–45.

Li, N., Loizou, P.C., 2008. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J. Acoust. Soc. Am. 123, 1673–1682.

McAulay, R.J., Quatieri, T.F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. Acoust. Speech Sig. Proc 744–754 ASSP-34.

Maclin, R., Opitz, D., 1997. An empirical evaluation of bagging and boosting. In: Proceedings of the 14th National Conference on Artifical Intelligence, Providence (1997).

Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2008. Phoneme representation and classification in primary auditory cortex. J. Acoust. Soc. Am. 123, 899–909.

Mitra, V., Franco, H., Graciarena, M., Mandal, A., 2012. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, pp. 4117–4120 March 25-30.

Moore, B.C.J., Glasberg, B.R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74, 750–753.

Moore, B.C.J., Vickers, D.A., Plack, C.J., Oxenham, A.J., 1999. Inter-relationship between different psychoacoustic measures assumed to be related to the cochlear active mechanism. J. Acoust. Soc. Am. 106, 2761–2778.

Moritz, N., Anemüller, J., Kollmeier, B., 2015. An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition. IEEE Trans. Audio Speech Lang. Proc. 23, 1926–1937.

Muralimanohar, R.Kumar, 2018. Analyzing the contribution of envelope modulations to the intelligibility of reverberant speech.

Naftaly, U., Intrator, N., Horn, D., 1997. Optimal ensemble averaging of neural networks. Netw.: Comput. Neural Syst. 8, 283–296.

Ng, E.H., Rudner, M., Lunner, T., Pedersen, M.S., Rönnberg, J., 2013. Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. Int. J. Audiol. 52, 433–441.

Nilsson, M., Soli, S.D., Sullivan, J., 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. 95, 1085–1099.

Opitz, D., Maclin, R., 1999. Popular ensemble methods: an empirical study. J. Artif. Intell. Res 11, 169–198.

Patterson, R.D., Allerhand, M.H., Giguère, C., 1995. Time-domain modeling of peripheral auditory processing: a modular architecture and a software platform. J. Acoust. Soc. Am. 98, 1890–1894.

Rasetshwane, D.M., Raybine, D.A., Kopun, J.G., Gorga, M.P., Neely, S.T., 2019. Influence of instantaneous compression on recognition of speech in noise with temporal dips. J. Am. Acad. Audiol. 30, 16–30.

Relaño-Iborra, H., May, T., Zaar, J., Scheidiger, C., Dau, T., 2016. Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain. J. Acoust. Soc. Am. 140, 2670–2679.

Rhebergen, K.S., Versfeld, N.J., 2005. A speech intelligibility index based approach to predict the speech reception threshold for sentences in fluctuating noise for normal–hearing listeners. J. Acoust. Soc. Amer. 117, 2181–2192.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Phil. Trans. R. Soc. Lond. B 336, 367–373.

Rothauser, E.H., 1969. IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio Electroacoustics 17, 225–246.

Rousselet, G.A. 2017. How to compare dependent correlations. Downloaded from https://garstats.wordpress.com/2017/03/01/comp2dcorr/. (Last viewed 28 January 2020).

Ruggero, M.A., Rich, N.C., Recio, A., Narayan, S., Robles, L., 1997. Basilar-membrane responses to tones at the base of the chinchilla cochlea. J. Acoust. Soc. Am. 101, 2151–2163.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D., McClelland, F. (Eds.), Parallel Distributed Processing Vol. 1. MIT Press, Cambridge Mass.

Salehi, H., Parsa, V., Folkeard, P., 2018. Electroacoustic assessment of wireless remote microphone systems. Audiol. Res. 8:204, 16–23.

Souza, P.E., Arehart, K.H., Kates, J.M., Croghan, N.B.H., Gehani, N., 2013. Exploring the limits of frequency lowering. J. Speech Lang. Hear. Res. 56, 1349–1363.

Steinmetzger, K., Zaar, J., Relaño-Iborra, H., Rosen, S., Dau, T., 2019. Predicting the effects of periodicity on the intelligibility of masked speech: An evaluation of different modelling approaches and their limitations. J. Acoust. Soc. Am. 146, 2562–2576.

Sullivan, G.M., Feinn, R., 2012. Using effect size – or why the *p* value is not enough. J. Grad. Med. Ed. 4, 279–282.

Suzuki, Y., Takeshima, H., 2004. Equal-loudness-level contours for pure tones. J. Acoust. Soc. Am. 116, 918–933.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49, 1255.

Tufts, J.B., Molis, M.R., 2007. Perception of roughness by listeners with sensorineural hearing loss. J. Acoust. Soc. Am. 121, EL161–EL167.

Van Kuyk, S., Kleijn, W.B., Hendriks, R.C., 2018. An evaluation of intrusive instrumental intelligibility metrics. IEEE/ACM Trans. Audio Speech Lang. Proc. 26, 2153–2166.

Wasserman, P. D., 1989. *Neural Computing: Theory and Practice* (Van Nostrand Reinhold, New York), pp. 43-59.

Werbos, P.J., 1990. Backpropagation through time: what it does and how to do it. Proc. IEEE 78, 1550–1560.

Wojtczak, M., Biem, J.A., Micheyl, C., Oxenham, A.J., 2012. Perception of across-frequency asynchrony and the role of cochlear delay. J. Acoust. Soc. Am. 131, 363–377.

Zahorian, S.A., Rothenberg, M., 1981. Principal-components analysis for low-redundancy encoding of speech spectra. J. Acoust. Soc. Am. 69, 832–845.

Zeng, F-G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y-Y., Chen, H., 2004. On the dichotomy in auditory perception between temporal envelope and fine structure cues. J. Acoust. Soc. Am. 116, 1351–1354.

Zhang, X., Heinz, M.G., Bruce, I.C., Carney, L.H., 2001. A phenomenological model for the response of auditory nerve fibers: I. Nonlinear tuning with compression and suppression. J. Acoust. Soc. Am. 109, 648–670.

Zio, E., 2006. A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. IEEE Trans. Nucl. Sci. 53, 1460–1478.