4pPP9. An auditory model for intelligibility and quality predictions

James Kates*

*Corresponding author's address: Speech Language and Hearing Sciences, University of Colorado Boulder, 409 UCB, Boulder, CO 80309, James.Kates@colorado.edu

The perceptual effects of audio processing in devices such as hearing aids can be predicted by comparing auditory model outputs for the processed signal to the model outputs for a clean reference signal. This paper presents an improved auditory model that can be used for both intelligibility and quality predictions. The model starts with a middle-ear filter, followed by a gammatone auditory filter bank. Two-tone suppression is provided by setting the bandwidth of the control filters wider than that of the associated analysis filters. The analysis filter bandwidths are increased in response to increasing signal intensity, and compensation is provided for the variation in group delay across the auditory filter bank. Temporal alignment is also built into the model to facilitate the comparison of the unprocessed reference with the hearing-aid processed signals. The amplitude of the analysis filter outputs is modified by outer hair-cell dynamic-range compression and inner-hair cell firing-rate adaptation. Hearing loss is incorporated into the model as a shift in auditory threshold, an increase in the analysis filter bandwidths, and a reduction in the dynamic-range compression ratio. The model outputs include both the signal envelope and scaled basilar-membrane vibration in each auditory filter band.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

Auditory models form the basis of many procedures for predicting speech intelligibility and quality. The use of an auditory model is based on the assumption that the accuracy of speech intelligibility and quality predictions will benefit from embedding an auditory model into the metric. The objective of the model in these applications is not to reproduce every aspect of auditory signal processing, but rather to reproduce important aspects of peripheral signal processing while maintaining a reasonable degree of computational efficiency. If the application involves hearing aids or impaired hearing, then peripheral hearing loss must also be an integral part of the model.

The simplest auditory model is a filter bank representing the frequency analysis of the human ear. Additional complexity can be added depending on the purpose of the model. The speech intelligibility index (SII), for example, incorporates an auditory filter bank, to which corrections are applied to account for frequency-domain masking, signal intensity, and shifts in the auditory threshold (French and Steinberg, 1947; ANSI S3.5, 1997).

An alternative to applying corrections to a filter-bank model is to base the model more directly on auditory physiology. Models based all or in part on physiology have been used for predicting intelligibility (Holube and Kollmeier, 1996; Elhilali et al., 2003; Zilany and Bruce, 2007; Christiansen et al., 2010; Taal et al., 2011) and for quality (Tan et al., 2004; Huber and Kollmeier, 2006; Tan and Moore, 2008; Kates and Arehart, 2010). Of the quality models that incorporate hearing loss and satisfy the requirement of computational efficiency, the Kates and Arehart (2010) model appears to be the most accurate.

The model presented in this paper is an extension of the Kates and Arehart (2010) auditory model. That model has been shown to give outputs that can be used to produce accurate predictions of speech quality for normal-hearing and hearing-impaired listeners under a wide variety of noise, nonlinear distortion, and linear filtering conditions. The improvements in the new model include ensuring that the filter characteristics are independent of the signal sampling rate, increasing the model bandwidth to better analyze music signals, adjusting the auditory filter bandwidth in response to the signal intensity, a more accurate representation of cochlear dynamic-range compression, the inclusion of inner hair cell firing-rate adaptation, and compensation for the group delay of the auditory filter bank.

AUDITORY MODEL

Model Overview

The model inputs are the unprocessed reference and processed signals that are to be compared. The processing can include linear filtering, nonlinear signal manipulations, nonlinear distortion, and background noise. The model outputs are the envelope and basilar membrane vibration of the reference and processed signals in auditory frequency bands. The overall model block diagram is presented in Fig 1. The model operates at 24 kHz, so the first processing step is the sample-rate conversion of the signals. The comparison of the processed and reference signals generally 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 signal features for comparing the processed signal with the reference signal. This step is part of the performance index rather than an inherent aspect of the auditory model and is not described in this paper.

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 effects of hearing loss 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 a function of the amount of 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 basilar membrane 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 vibration tracks the envelope output. The auditory threshold for the vibration signal is represented as a low-level additive white noise.
Sample Rate Conversion

The middle-ear filter and the filters in the auditory filter bank are all infinite impulse response (IIR) designs. The magnitude and phase response of an IIR filter depends on the sampling rate, so filters having the same design specifications will differ if the sampling rates differ. Therefore, to ensure identical filter behavior for all input signals, the signals are resampled at 24 kHz. This sampling rate was chosen to minimize computational requirements while still providing adequate bandwidth for the highest frequency band (8 kHz) used for the auditory analysis.

Middle Ear

The primary purpose of the middle ear model is to reproduce the low-frequency and high-frequency attenuation observed in the equal-loudness contours at low signal levels (Suzuki and Takeshima, 2004). The low-frequency attenuation is represented by a 2-pole IIR high-pass filter having a cutoff frequency of 350 Hz. The high-frequency attenuation is represented by a 1-pole IIR low-pass filter having a cutoff frequency of 5 kHz (Kates, 1991).

Analysis Filter Bank

The parallel filter bank used for the auditory analysis consists of fourth-order gammatone filters (Patterson et al., 1995). The digital gammatone filter bank is implemented using the base-band impulse-invariant method (Cooke, 1991; Immerseel and Peeters, 2003). A total of 32 filters are used to cover the frequency range of 80 to 8000 Hz. This frequency range is greater than the 150 to 8000 Hz used in the SII standard (ANSI S3.5, 1997), and was chosen to accommodate music as well as speech signals. A linear filter bank is used for computational efficiency. The linear filter bank leaves out the dynamic interaction between the instantaneous signal level and the cochlear filter shape (Zhang et al., 2001), but the 24-kHz sampling rate gives a significant computational savings over the 500-kHz sampling rate required for the adaptive-filter (Zhang et al., 2001) model.

Outer Hair-Cell Damage

Hearing loss can be caused by both damage to the outer hair cells that control the cochlear filters and by damage to the inner hair cells that perform the mechanical-to-neural transduction (Liberman and Dodds, 1984). OHC damage is modeled as a reduction in the quality factor (Q) of the cochlear filters, resulting in increased filter bandwidth and reduced gain (Kates, 1991). IHC damage is modeled as a reduction in the sensitivity of the neural transduction mechanism. For moderate hearing losses, approximately 80 percent of the total loss given by the audiogram can be ascribed to OHC damage (Moore et al., 1999), with the remainder ascribed to IHC damage.

Hearing loss is incorporated into the gammatone filter bank as an increase in filter bandwidth. The data of Moore et al. (1999) indicate that over approximately the first 50 dB of the total hearing loss there is a strong correlation of the auditory filter bandwidth with loss; the filter bandwidth increases by about a factor of two over this range. As the total loss increases beyond 50 dB, however, the filter bandwidth increases rapidly. As an approximation to this
behavior, the filter bandwidth relative to that of a normal ear is given by $BW = 1 + (\text{attn}/50) + 2(\text{attn}/50)^6$, where \( \text{attn} \) is the hearing loss in dB ascribed to the OHC damage, with a maximum attenuation of 50 dB.

The responses of the gammatone filters are normalized so that the gain for a signal at the filter center frequency is 0 dB. The reduction of the filter gain caused by the OHC loss is implemented in the compression stage that follows the filter bank. The filter transfer functions for normal hearing are plotted in Fig 3 for the filter center frequency range of 80 Hz to 8 kHz. The filter shapes for the maximum OHC loss allowed in the model as a function of frequency are plotted in Fig 4. The OHC damage causes a broadening of the filter around its center frequency as well as an increase in the filter response at low frequencies (Liberman and Dodds, 1984).

**FIGURE 3.** Gammatone filter magnitude transfer functions for normal hearing.

**FIGURE 4.** Gammatone filter magnitude transfer functions for maximum hearing loss and the control filters.

### Signal Intensity

The shape of the auditory filters depends on the intensity of the input signal as well as on the degree of hearing loss. Measurements of the basilar-membrane vibration in animals (Rhode, 1971; Ruggero et al., 1997) show that the auditory filters become broader as the signal level increases. Behavioral measurements of auditory filter nonlinearity have also been made in humans (Glasberg and Moore, 2000; Baker and Rosen, 2002; Baker and Rosen, 2006). In contrast to the animal studies, the human data tend to show nearly constant filter bandwidths for intensities below 50 dB SPL for both normal-hearing (Baker and Rosen, 2006) and hearing-impaired (Baker and Rosen, 2002) listeners. The bandwidths increase with increasing intensity above 50 dB SPL, and a linear function is used in this paper to approximate the increase in bandwidth with intensity.

An example of the linear approximation is shown schematically in Fig 5. For normal hearing, the filter bandwidth is set to the ERB (Moore and Glasberg, 1983) for intensities below 50 dB SPL. For impaired hearing, the bandwidth at and below 50 dB SPL is set to the bandwidth computed for the amount of OHC damage related to the hearing loss. For both normal and impaired hearing, the bandwidth is set to the bandwidth corresponding to maximum OHC damage for intensities at or above 100 dB SPL. Linear interpolation is used for intensities between 50 and 100 dB SPL. In the limiting case of a hearing loss giving the maximum amount of OHC damage, the bandwidth stays at the maximum value at all signal levels. The filter bandwidth is determined by the signal intensity in the control filter bank outputs and remains constant throughout the auditory analysis filtering operation.

### Control Filter Bank

Cochlear mechanics provides nearly instantaneous dynamic-range compression. In the cochlea the gain changes are combined with dynamic changes in the filter bandwidth, e.g. the filter Q is dynamically varied in response to the signal level (Zhang et al., 2001). In the simplified cochlear model used here, the compression is a separate stage that follows the linear filter bank. The compression gain is computed using the envelope in each band of the control filter bank, and the compression gain multiplies the signal in each auditory analysis filter band.

The control filter bank, like the analysis filter bank, uses fourth-order gammatone filters. The filter bandwidths for the control filters are set to correspond to the maximum bandwidth allowed in the model, as shown in Fig 4. The control filter bandwidths thus match the auditory analysis bandwidths for the maximum hearing loss, and are wider than the auditory analysis filters for reduced hearing loss and normal hearing. The wide control filters provide two-tone suppression in the auditory model (Zhang et al., 2001; Heinz et al., 2001; Bruce et al., 2003). The center
frequency of each control filter is shifted higher in frequency relative to the corresponding auditory analysis filter. The frequency shift corresponds to a fractional basal shift of 0.02 of the length of the cochlear partition using a human frequency-position function (Greenwood, 1990). This frequency shift is less than the basal shift used by Zhang et al. (2001) to model the cat cochlea; however, the shift produces two-tone suppression results that are consistent with the human psychophysical measurements recorded by Duifuis (1980) for a probe at 50 dB SPL.

**Figure 5.** Example of the adjustment of the auditory filter bandwidth with increasing signal intensity.

**FIGURE 6.** Input/output relationship showing the dynamic-range compression due to OHC function.

### Dynamic-Range Compression

The control signal envelope is the input to the compression rule. The compression gain is then passed through an 800-Hz low-pass filter to approximate the compression time delay observed in the cochlea (Zhang et al., 2001). The compression rule for normal hearing is modeled by three line segments as shown by the bold lines in Fig 6. Inputs within 30 dB of normal auditory threshold (0 dB SPL) receive linear gain. Inputs between 30 and 100 dB SPL are compressed. The system reverts to linear gain for inputs above 100 dB SPL. The compression ratio in the model for normal hearing increases linearly with ERB number from a compression ratio of 1.25:1 at 80 Hz to a compression ratio of 3.5:1 at 8 kHz. This compression behavior is consistent with physiological measurements of compression in the cochlea (Cooper and Rhode, 1997; Lopez-Poveda and Alves-Pinto, 2008) and with psychophysical estimates of compression in the human ear (Hicks and Bacon, 1999; Plack and Oxenham, 2000).

OHC damage shifts the auditory threshold and reduces the compression ratio, as shown by the thin lines in Fig 6. The dependence of the compression behavior on OHC damage reproduces the changes in the auditory-nerve firing rate measured in damaged cochleas (Heinz et al., 2005; Neely et al., 2009) and the loudness recruitment found in hearing-impaired listeners (Ksiessling, 1993). The shifted curves are constructed so that an input of 100 dB SPL in a given frequency band always produces the same output level independent of the amount of OHC damage. The maximum gain reduction $D$ shown in the figure due to OHC damage is a function of the normal-hearing compression ratio in the frequency band. The maximum gain reduction is 14 dB for the compression ratio of 1.25:1 at 80 Hz, and increases to 50 dB for the compression ratio of 3.5:1 at 8 kHz.

In each frequency band, the OHC threshold is set to $1.25D$. If the total hearing loss given is greater than this threshold, the OHC loss is set to $D$ and the remaining loss is ascribed to IHC damage. For this condition the compression system is reduced to linear amplification as shown by the line having the $x$-axis intercept at $D$ in Fig 6. If the total hearing loss is less than the threshold, 80 percent of the loss is ascribed to OHC damage and 20 percent to IHC damage. This condition results in a reduction in the OHC gain of less than $D$ combined with a compression ratio partway between 1:1 and maximum compression. This behavior is shown in Fig 6 for the line having the $x$-axis intercept at $d$. The lower compression knee point is always set to 30 dB above the auditory threshold, which is a change from the Kates and Arehart (2010) model.

### Two-Tone Suppression

In two-tone suppression, the presence of a signal outside the bandwidth of the analysis filter reduces the response to a probe signal near the center frequency of the analysis filter (Sachs and Kiang, 1968; Duifuis, 1980; Delgutte,
Two-tone suppression is greatest in the normal ear, and is substantially reduced or eliminated in the impaired ear (Schmiedt, 1984). The compression input-output function, when combined with the control filter bank, produces two-tone suppression in the cochlear model. The control filter is wider than the corresponding analysis filter, which allows the presence of a signal outside the bandwidth of the analysis filter but still within the bandwidth of the control filter to reduce the gain for a signal within the analysis filter passband.

Two-tone suppression in the cochlear model is illustrated in Fig. 7 for normal hearing. The probe in this example is a 50-dB SPL sinusoid at a frequency of 2080 Hz, which is the center frequency for band 20 of the 32 analysis bands. Suppressor tones outside the set of contours reduce the compressed signal output by less than 1 dB compared to the output for the probe alone. A tone at 90 dB SPL at a frequency of 900 Hz, on the other hand, will reduce the output within the probe frequency band by an additional 9 dB. Suppression is reduced in impaired hearing because the analysis filter bandwidth is increased and the compression ratio is reduced with increasing hearing loss. In the limit of maximum OHC damage, the analysis and control filters have equal bandwidths and the auditory compression becomes a linear system, thus completely eliminating the two-tone suppression in the model.

**Temporal Alignment**

The model contains three stages of temporal alignment of the processed signal with the reference signal as shown in Fig. 1. The first stage occurs at the input to the model; in this stage the signals are approximately aligned and the signal durations matched. The alignment is based on the maximum of the broadband signal cross-correlation. A second, band-by-band alignment occurs after the gammatone filter bank frequency analysis and the dynamic-range compression. The envelope and basilar-membrane vibration outputs for the processed signal are separately matched to the reference signal. The match is based on the maximum of the cross-correlation of the signals. As a result of this temporal alignment, the processed signal has the same delay as a function of frequency as the reference signal, and the group delay associated with the hearing-aid or other audio processing is removed.

The group delay of the gammatone filters is a good match to the latency measured in human ears (Don et al., 1998). However, there is evidence that compensation for the frequency-dependent group delay occurs higher in the auditory pathway (Uppenkamp et al., 2001; Wojtczak et al., 2012). Delay compensation has therefore been provided for the auditory model output. The group delay for the reference signal at the center frequency of each band is computed, and delay is added to each band of the reference and processed signals so that the total group delay in each band matches that of the lowest-frequency band.

**dB Conversion**

The envelope signal, after dynamic-range compression, is converted to dB above auditory threshold. Normal threshold is used since attenuation due to OHC damage has already been applied to the signals. The hearing loss due to IHC damage is applied as an additional attenuation after the dB SL conversion. The basilar membrane vibration signal is multiplied by the same gain factor as computed for the envelope dB conversion so that the vibration signal amplitude tracks the dB envelope. The compressed average outputs in dB SL correspond to firing rates in the auditory nerve (Sachs and Abbas, 1974; Yates et al., 1990) averaged over the population of inner hair-cell synapses.
Inner Hair-Cell Synapse

The IHC synapse provides the rapid and short-term adaptation observed in the neural firing rate (Harris and Dallos, 1979; Gorga and Abbas, 1981). The synapse is a simplified two-reservoir model based on the models of Westerman and Smith (1987) and Kates (1991). The rapid adaptation time constant is 2 ms and the short-term time constant is 60 ms. The adaptation emphasizes sudden changes in the signal level, such as occurs at the onset of a stop consonant. The model is adjusted so that an instantaneous jump of 20 dB in the input signal level produces a peak output 20 dB above the steady-state response. The adaptation is computed for the envelope signal in dB SL, and then the basilar-membrane vibration signal is multiplied by the same gain-versus-time function. Hearing loss due to IHC damage is implemented as an attenuation of the signal at the input to the synapse model. The differential equations that describe the analog circuit were transformed into the digital domain using first-order backwards differences in a state-space representation at the 24-kHz sampling rate.

An example of the synapse response is shown in Fig 8. The compressed signal envelope in the frequency band is initially at 40 dB SL. The level jumps to 60 dB SL at 100 ms, and returns to 40 dB SL at 600 ms. Raised-cosine 5-ms windows are applied to both level changes, which reduces the overshoot in comparison with instantaneous transitions. The peak of the transient response at 100 ms is 71.4 dB, and the minimum at 600 ms is 28.6 dB SL.

Long-Term Average

Some models of intelligibility and quality make use of the signal long-term average spectrum (French and Steinberg, 1947; Theide et al., 2000; Beerends et al., 2002; Moore and Tan, 2004; Kates and Arehart, 2010), as do models of loudness (Moore and Glasberg, 2004; Chen et al., 2011). A set of long-term average signals is therefore provided as an additional auditory model output.

The root-mean-squared (RMS) average output is computed for the envelopes in each of the auditory analysis filters after the filter bandwidths have been adjusted for OHC damage and signal intensity. The RMS average outputs are also computed for the control filter bank signals. The average control signal is converted to dB above threshold, and the input/output compression rule as shown in Fig 8 is used to compute compression gains for the averaged signals as a function of frequency. The compression gain in dB is then added to the dB levels in each of the analysis filter bands, after which the attenuation due to IHC damage is applied to the average signals in each band.

As for the envelope described in Section 2.9, the compressed average outputs in dB SL correspond to average firing rates in the auditory nerve (Sachs and Abbas, 1974; Yates et al., 1990). The cited loudness models, on the other hand, use the specific loudness in each frequency band. The specific loudness is proportional to the mean-square level in each band raised to the 0.2 power (Moore and Glasberg, 2004). In the auditory model, taking the RMS signal level raises the mean-square level to a power of 0.5, and the dynamic-range compression of 2.5:1 at mid frequencies further reduces the power to the vicinity of 0.2. Thus specific loudness, to within a scale factor, can be approximated from the average auditory model outputs by converting the dB SL values to linear amplitude.

DISCUSSION AND CONCLUSIONS

The auditory model presented in this paper is designed to be the initial processing stage for intelligibility and quality indices. Because it is intended for practical applications, computational efficiency is an important aspect of the implementation. The goal is to efficiently approximate the salient auditory behavior rather than provide an exact but potentially time-consuming model.

A significant computational savings, for example, is realized by using linear filters for the auditory analysis rather than trying to duplicate the instantaneous filter gain and bandwidth changes mediated by the outer hair cells. The auditory filter bandwidth is determined by the hearing loss and the average signal intensity in the control filters prior to sending the signal through the analysis filters, thus allowing for efficient analysis filters having constant bandwidth. A further justification for this approach is that intelligibility and quality predictions generally use short speech sequences presented at conversational levels, so large variations in signal intensity would not be expected.

The IHC synapse model is also greatly simplified in comparison with the ear. The auditory model does not produce a neural spike-train output. It merely approximates the firing rate rapid and short-term adaptation that impacts the neural firing patterns. The outputs are the envelope and a vibration signal in each analysis band that has the same envelope but also contains the temporal fine structure. The vibration signal has not been rectified; it remains zero-mean to facilitate computing the band-by-band cross-correlations between the reference and processed signals that are used in some intelligibility (Kates and Arehart, 2005) and quality (Tan et al., 2004) indices.
The auditory model processes the clean reference and the hearing-aid output signals as a set of combined inputs. The reason for the combined processing is the temporal alignment of the two signals, which is needed for the subsequent intelligibility or quality index calculations. The processing delay associated with the hearing-aid output is removed, and compensation is also provided for the group delay associated with the auditory analysis filters.

The choice of reference signal processing in the auditory model is tied to whether the model outputs will be used for intelligibility or quality. Intelligibility is measured on an absolute scale, typically phonemes, words, or sentences correct. A general assumption is that the hearing-impaired ear cannot produce higher recognition scores than a normal-hearing ear. Therefore, if the auditory model output is intended for intelligibility, the reference is the clean signal processed using a model of normal hearing while the hearing-aid output is processed through a model of the impaired ear. Quality ratings, on the other hand, are assumed to be based on comparing the hearing-aid output to the best signal quality that can be perceived by the hearing-impaired listener (Kates and Arehart, 2010). The reference signal for a quality index should therefore be the clean input to the hearing aid combined with linear amplification (e.g. Byrne and Dillon, 1986) to compensate for the hearing loss, and this amplified reference is processed through a model of the impaired ear.

REFERENCES


