INVITED REVIEW

Computational models for predicting sound quality

Brian C. J. Moore*

*Department of Experimental Psychology, University of Cambridge,
Downing Street, Cambridge, CB2 3EB, United Kingdom*

Abstract: The quality of an audio device depends on how accurately the device transmits the properties of the sound source to the ear(s) of the listener. Two types of “distortion” can occur: (1) “Linear” distortion, namely a deviation of the frequency response from the “target” response; (2) Nonlinear distortion, which is characterised by frequency components in the output of the device that were not present in the input. These two forms of distortion have different perceptual effects. Their effects on sound quality can be predicted using a model of auditory processing with the following stages: filters simulating the transmission of sound from the device to the ear of the listener and through the middle ear; and an array of bandpass filters simulating the filters that exist in the cochlea. For predicting the perceptual effects of linear distortion, a model operating in the frequency domain can be used. For predicting the perceptual effects of nonlinear distortion, a model operating in the time domain is required, since the detailed waveforms at the outputs of the auditory filters need to be considered. The models described give accurate predictions for a wide range of “artificial” and “real” linear and nonlinear distortions.

Keywords: Sound quality, Distortion, Auditory models, Excitation patterns

PACS number: 43.66.Ba, 43.66.Lj [doi:10.1250/ast.41.75]

1. INTRODUCTION

The perceived quality of sounds reproduced by devices such as loudspeakers, headphones and telephones depends on how accurately the sounds are reproduced. Inaccuracies in sound reproduction are often described as “distortion.” There are two main types of distortion: (1) “Linear” distortion, which is associated with the imperfect amplitude and phase response of the device; (2) “Nonlinear” distortion, which is associated with frequency components in the output of the device that were not present in the input. These have different perceptual effects. Linear distortion is often described as a “colouration” of the sound or a change in timbre, and is characterised by adjectives such as sharp, bright, dull, boomy, and nasal. In principle, the effects of linear distortion can be compensated by a linear “inverse” filter. Nonlinear distortion is described by adjectives such as gritty, harsh, rough, and noisy. Once nonlinear distortion has been introduced, it is difficult or impossible to remove it using simple signal processing.

This paper describes a series of studies that were conducted in collaboration with Chin-Tuan Tan to characterise and model the perceptual effects of linear and

nonlinear distortion and their combination. Both artificially distorted signals and naturally distorted signals were used.

2. CHARACTERISING LINEAR DISTORTION

Moore and Tan [1] conducted a study to explore how the perceived naturalness (assumed to be closely related to sound quality) of music and speech signals was affected by various forms of digital linear filtering. Some of the types of filtering were intended to mimic the spectral distortions that can occur in transducers. These included spectral tilts and band limiting. Regular spectral ripples were also used. The ripples were uniformly spaced on an ERB_N -number scale, which is a perceptually relevant transformation of the frequency scale with units Cams [2]. The equation relating ERB_N number to frequency, f (hertz), is:

$$ERB_N \text{ number (Cams)} = 21.4 \log_{10}(0.00437f + 1) \quad (1)$$

Some of the filter characteristics were deliberately designed to give rather extreme degradations of perceived naturalness, so as to provide listeners with clear examples of unnatural sound qualities.

The sounds were reproduced by Sennheiser HD580 headphones, which have a good approximation to a diffuse-field characteristic. They were selected for their low nonlinear distortion. This ensured that the main factor

*e-mail: bcjm@cam.ac.uk

Fig. 1 Examples of the frequency responses used by Moore and Tan [1].

affecting the subjective responses was the spectral filtering. Examples of the filter characteristics used are given in Fig. 1.

The overall level of each filtered signal was adjusted so as to give a constant loudness level of 86.4 phons (for binaural listening), calculated using the model of Moore *et al.* [3]. In a given session, a subject was tested using either speech or music stimuli. Filtered stimuli were presented in a randomized order. After each stimulus presentation, the subject was required to rate the perceived quality on a 10-point scale where “10” indicates “very natural - uncolored” and “1” represents “very unnatural - highly colored.”

The results were broadly similar for speech and music, and for brevity only the results for music are presented. Figure 2 shows the mean ratings for music for stimuli with rather large spectral ripples (15-dB peak-to-valley ratio) with ripples densities from 0.05 to 0.5 ripples/Cam, extending over various frequency ranges. The ripples produced the greatest degradation of sound quality when they were dense (0.2 and 0.5 ripples/Cam) and extended over a wide frequency range (87–6,981 Hz).

Figure 3 shows the mean ratings for stimuli with spectral tilts. Large negative or positive tilts (± 1 dB/Cam) led to large degradations in sound quality, especially when the tilts extended over a wide frequency range.

Fig. 2 Mean ratings for music stimuli with spectral ripples.

Fig. 3 Mean ratings for music stimuli with spectral tilts.

Fig. 4 Mean ratings for bandlimited music stimuli.

Figure 4 shows the mean ratings for bandlimited stimuli. Reducing the upper cutoff frequency from 16,854 to 3,547 or increasing the lower cutoff frequency from 55

to 313 Hz degraded sound quality. The lowest quality was obtained for the most severe bandlimiting (313–3,547 Hz), which corresponds to the typical telephone frequency range. Sound quality was also very low for the frequency range typically covered by hearing aids (208 to 5,583 Hz) [4].

3. MODELLING THE EFFECTS OF LINEAR DISTORTION

3.1. Calculation of the Excitation Pattern

The excitation pattern of a sound is a representation in the auditory system of the short-term spectrum of the sound; it resembles a blurred version of that spectrum [5]. The excitation pattern is defined as the output level of the auditory filters plotted as a function of filter centre frequency [6]. Each auditory filter represents the “tuning” of the auditory system at one centre frequency [7,8]. It is assumed that changes in the excitation pattern produced by the linear distortion reduce perceived naturalness.

The characteristics of the auditory filters have usually been estimated using simultaneous masking, often using the notched-noise method [8,9]. There is evidence that auditory filter shapes measured in this way are broader than the underlying filter shapes that determine the excitation pattern [10–12]. In what follows, excitation patterns were calculated using methods similar to those described in Glasberg and Moore [8] and Moore *et al.* [3], but the overall sharpness of the auditory filters was adjusted to give the best fit to the data. The sharpness of the filters at each centre frequency is controlled by a parameter p ; the higher the value of p , the sharper is the filter. The sharpening was implemented by multiplying the value of p at each centre frequency by a frequency-independent factor, s .

It was assumed that differences in excitation patterns for steady sounds, produced by a given linear distortion, reflect the perceived change in naturalness induced by that distortion for real-life time-varying signals. Excitation patterns were calculated for two steady input signals, a noise with the same long-term average spectrum as speech [13], and a broadband pink noise, which is reasonably representative of some types of music signals. Excitation patterns were calculated for each original (undistorted) input signal and for that signal after it had been passed through each of the filters described by Moore and Tan [1]. The input level of the stimuli was as used for stimulus presentation by Moore and Tan [1]. The calculated excitation patterns took into account the diffuse-field response of the Sennheiser HD580 earphones that were used by Moore and Tan [1].

3.2. Quantifying Excitation-pattern Differences

Excitation patterns were calculated on an ERB_N -

number scale. The maximum ERB_N number, i , was 40 Cams, corresponding to a frequency of about 16,800 Hz. When the calculated excitation level for any i fell below a “floor” value, f_i , the excitation level was set equal to f_i .

To derive measures of the differences between the excitation patterns of the original and distorted signals, the patterns were sampled at 0.5-Cam intervals. Denote the excitation level evoked by the original signal at the i th Cam value by $EO(i)$ and the corresponding excitation level evoked by the distorted signal $ED(i)$. Two measures of the overall difference were initially evaluated, but I describe here only the measure that was found to give the most accurate predictions of sound quality, which was the standard deviation (SD) of the values of $EO(i) - ED(i)$ across all ERB_N numbers.

Comparison of the magnitude of this measure with the naturalness ratings indicated an orderly relationship for spectral tilt distortions, but the relationship was much less orderly for spectral ripple distortions. The reason for this can be understood by consideration of the data in Fig. 2. These show that perceived quality decreased with increasing ripple density (rate). For a given ripple depth, closely spaced ripples degraded quality more than widely spaced ripples, except when the ripple rate was above 0.2 ripples/Cam.

The results suggest that quality ratings are influenced by the rate of change over frequency of excitation-pattern differences produced by the spectral distortion. This was quantified using the “second-order” excitation-pattern difference, defined by $[EO(i+1) - EO(i)] - [ED(i+1) - ED(i)]$. This corresponds to the difference in local slope of the excitation patterns for the original and distorted signals. The overall second-order difference was calculated as the SD of the values of $[EO(i+1) - EO(i)] - [ED(i+1) - ED(i)]$ across all ERB_N numbers.

The results of Moore and Tan [1] indicated that spectral distortions at high frequencies had a slightly smaller perceptual effect than spectral distortions at lower frequencies. Also, the results for the band-limiting conditions indicated relatively small effects of removing the highest frequencies in the signals. To accommodate this, a weighting function was used that reduced the contribution of high frequencies to the overall excitation difference. This weighting function was flat for ERB_N numbers up to 17 Cams (corresponding to about 1,200 Hz) and decreased with increasing ERB_N number above that. The relative weighting, $W(i)$, as a function of ERB number, i , was thus:

$$W(i) = 1 \text{ for } i < 17.5 \quad (2)$$

$$W(i) = 1 - w_s(i - 17.0)/46 \text{ for } i \text{ from } 17.5 \text{ to } 40 \quad (3)$$

The slope of the decreasing part of the weighting function, w_s , was a free parameter of the model. The same

weighting was applied to the first-order differences and the second-order differences.

In summary the model had four free parameters: the factor by which the simulated auditory filters were sharpened, s ; the relative weighting of the first- and second order excitation pattern differences, w ; the floor level, f ; and the across-frequency weighting, w_s .

The final stages of the model were: calculation of the SD of the weighted first-order differences across all i ; calculation of the SD of the weighted second-order differences across all i ; formation of a weighted sum of the SDs of the first-order and second-order differences, with weight parameter w . The final calculated quantity is called the overall weighted excitation-pattern difference D . It was assumed that subjective naturalness was inversely related to D and predictable from D .

Initial analyses showed that the relationship between subjective ratings, S , and D was curvilinear. To make the relationship more linear, the first step was to apply a linear transform to each value of S so that the range of transformed scores, S_t , was from 0 to 1:

$$S_t = (S - \text{MIN})/(\text{MAX} - \text{MIN}), \quad (4)$$

where MIN is the lowest obtained rating and MAX is the highest.

The second step was to apply the arcsine transform to the values of S_t , giving the transformed values, T :

$$T = 2 \arcsin \sqrt{S_t} \quad (5)$$

The resulting T values were almost linearly related to the D values. Straight lines were fitted to scatter plots of T versus D , using a least-squares procedure. Lines were fitted separately to the data for music and speech stimuli, as the

values of D were, on average, lower for the speech stimuli, owing to the de-weighting of excitation differences at very low and high frequencies. The fitted lines were used to predict T values from the calculated D values for each of the conditions used by Moore and Tan [1]. The predicted T values were then transformed to predicted naturalness ratings, by inverting the operations defined by Eqs. (4) and (5).

3.3. Comparison of Obtained and Predicted Ratings

Scatter plots of obtained versus predicted rating for music and speech are shown in Fig. 5. There was an excellent correspondence between the two. The correlation between obtained and predicted ratings was 0.955 for both music and speech.

4. CHARACTERISING NONLINEAR DISTORTION

Tan *et al.* [14] assessed the effect of various types of nonlinear distortion on the perceived quality of speech and music signals. Some of the distortions, such as “hard” and “soft” peak clipping and centre clipping, were meant to resemble distortions that occur in electro-acoustical systems such as loudspeakers and amplifiers. Other distortions were not meant to mimic any physical system, but were used to provide a greater variety of types of distortion. An example is “full range” waveform distortion produced by raising the instantaneous absolute value of the waveform to a power α ($\neq 1$) while preserving the sign. Subjects were asked to rate the perceived amount of distortion on a ten-point scale (where 1 was most distorted and 10 least distorted). In one experiment, the distortions were applied to the broadband signals. In another experiment, the

Fig. 5 Scatter plot of obtained versus predicted sound quality ratings for music (left) and speech (right).

Fig. 6 Mean ratings for speech signals with artificial distortions applied to the broadband signal.

distortions were applied to sub-bands of the signal (0 to 606 Hz, 606 to 1,973 Hz, 1,973 to 5,583 Hz, and 5,583 to 22,050 Hz).

The results for speech for the distortions applied to the broadband signal are shown in Fig. 6. A similar pattern of results was obtained for music. The lowest ratings were obtained for the full-range modification with $\alpha = 0.5$ or 2; the mean ratings were close to 1 for these cases. Hard symmetrical clipping 10% of the time also gave low ratings, about 3.5 for speech and 3.1 for music. Soft clipping and centre clipping up to 2% had little effect on quality.

The results when the distortion was applied to sub-bands of the signal (not shown) indicated that the distortion in each sub-band made a roughly equal contribution to the overall perceived distortion.

5. MODELLING THE EFFECTS OF NONLINEAR DISTORTION

Tan *et al.* [15] proposed the following method for predicting the perceptual effects of nonlinear distortion. When nonlinearities are present in a system, the output of the system in a given frequency region may be influenced by input components in a different frequency region. These regions were characterised using an array of simulated auditory filters [2,16] covering the audible frequency range. The amount of distortion at the output of a given simulated

Fig. 7 Block diagram of the model for predicting the perceptual effects of nonlinear distortion.

filter can be estimated from the correlation of the output of that filter in response to the original and distorted signals; the higher the correlation, the lower is the distortion, and the higher is the perceptual quality.

A block diagram of the model is shown in Fig. 7. The model uses a sample rate of 44.1 kHz and operates in the following way.

The input signal is speech (for prediction of perceptual judgments of speech) or music (for prediction of perceptual judgments of music). The model requires the waveform of the original input signal and the waveform of the signal after passing through the nonlinear system under test. The original and distorted signals are time aligned (not shown).

The original and distorted signals are filtered to mimic the effects of transmission through the outer and middle ear. This is done using a finite impulse response (FIR) filter with 4,097 coefficients, as described by Glasberg and Moore [17]. This filtering results in a relative attenuation of frequency components below 500 Hz and above 5,000 Hz.

The filtered signals are fed (separately) to an array of 40 gammatone filters each with a bandwidth of one ERB_N . Gammatone filters provide a computationally efficient time-domain simulation of the auditory filters [18]. The filter centre frequencies are spaced at one- ERB_N intervals, and cover the range from 50 Hz to 19,739 Hz (it is assumed that the lowest auditory filter is centred at 50 Hz [19]).

For each filter, the output is split into 30-ms non-overlapping frames, with index i . The cross-correlation is

calculated between the waveform of the i th frame at the filter output in response to the distorted signal and the waveform resulting from concatenation of frames $i - 1$, i , and $i + 1$ in response to the original signal, for lags from -10 to $+10$ ms. The cross-correlation for the j th filter is given as a function of lag index η (in samples) by:

$$r_{xy}(i; \eta; j) = \frac{\sum_{n=(i-1)L+1+\eta}^{iL+\eta} x(n; j)y(n - \eta; j)}{\sqrt{\left(\sum_{n=(i-1)L+1+\eta}^{iL+\eta} x(n; j)^2\right)\left(\sum_{n=(i-1)L+1+\eta}^{iL+\eta} y(n - \eta; j)^2\right)}} \quad (6)$$

where $\eta = -441$ to 441 .

For each frame and each filter, the maximum value of the normalized cross-correlation over all values of η , $X_{\max}$, is determined; this allows for any (small) frequency-dependent time delay in the nonlinear system. The value of $X_{\max}$ provides a measure of the extent to which the output in response to the distorted signal was affected by frequency components that were not present in the original signal; the lower $X_{\max}$, the more is the influence of distortion components. If $X_{\max}$ is equal to 1, this indicates no distortion.

For filters where the output is very low, the perceptual effect of distortion is likely to be small. A low output usually occurs at very low and high frequencies, because of the filter simulating the effect of the outer and middle ear. A low output might also occur because of the limited frequency range of the transducer that introduced the nonlinearity. To account for level effects, for each 30-ms frame, i , the power at the output of each filter, j , is calculated and converted to decibels according to:

$$\text{Level}(i, j) = 10 \log_{10} \left(\frac{1}{L} \sum_{n=(i-1)L+1}^{iL} y(n; j)^2 \right) \quad (7)$$

The value of $\text{Level}(i, j)$ is used to determine a weight that is applied to the value of $X_{\max}$ for frame i and filter j . The weights are scaled separately for each frame so that the sum of the weights is unity. The weighting function was determined by trying a large variety of functions and assessing how well the model predicted the data from the experiments of Tan *et al.* [14]. For each frame, the weighted values of $X_{\max}$ are summed across filters, to give an overall measure of correlation for that frame. The average correlation determined in this way is then averaged across frames. This gives a measure, R_{nonlin} , which decreases with increasing perceived distortion.

The relationship between mean subjective ratings, S , and the measure of distortion R_{nonlin} is somewhat curvilinear. The relationship was characterised with the following equation:

Fig. 8 Scatter plot of obtained versus predicted sound quality ratings for speech signals with broadband nonlinear distortions applied.

$$S = (a + bR_{\text{nonlin}}^c)/(a + R_{\text{nonlin}}^c) \quad (8)$$

The parameters, a , b and c were fitted to the subjective ratings separately for speech and music stimuli. The fitted functions were used to predict the mean subjective rating for a given type of distortion from the R_{nonlin} value for that distortion.

Figure 8 shows a scatter plot of obtained versus predicted values of S for various types of nonlinear distortion applied to broadband speech signals (mean ratings for these distorted signals are shown in Fig. 6). The correlation between obtained and predicted ratings was 0.98. The corresponding correlation for music signals was also 0.98.

6. CHARACTERISING AND MODELING THE PERCEPTUAL EFFECTS OF COMBINED LINEAR AND NONLINEAR DISTORTION

Moore *et al.* [20] studied the perceptual effects of combined linear and nonlinear distortion. In one experiment artificial distortions were used, applied either to the broadband signal or to limited frequency ranges. The distortions included: (1) spectral ripples combined with hard clipping; (2) spectral ripples combined with full-range distortion; (3) spectral tilts combined with hard clipping; (4) spectral tilts combined with full-range distortion.

To model the effects of these combined distortions, the models for predicting the effects of linear and nonlinear distortion, as described above, were initially run independently. The subjective ratings predicted by the linear model are denoted S_{lin} and the subjective ratings predicted by the

Fig. 9 Scatter plot of obtained versus predicted ratings for music stimuli subjected to various forms of combined linear and nonlinear artificial distortions.

Fig. 10 Scatter plot of obtained versus predicted ratings for music stimuli subjected to various forms of combined linear and nonlinear artificial and real distortions.

nonlinear model are denoted S_{nonlin} . The ratings were combined according to the rule:

$$S_{overall} = \beta S_{lin} + (1 - \beta) S_{nonlin} \quad (9)$$

where β is a parameter that characterizes the relative perceptual importance of the linear and nonlinear parts of the distortion. A good fit to the data was obtained with $\beta = 0.3$, indicating greater effects for nonlinear than for linear distortion.

Figure 9 is a scatter plot of obtained versus predicted ratings for music stimuli with the combined distortions described above. The correlation between obtained and predicted ratings was 0.95. The corresponding correlation for speech signals with combined distortions was 0.89.

In another experiment, Moore *et al.* [20] obtained ratings of sound quality for music and speech stimuli subjected to combined artificial distortions, a subset of those described above, and for stimuli with “real” distortions. These were produced using four compact electrodynamic transducers (13–20 mm in diameter) mounted in a variety of acoustic structures, selected to be representative of earpieces often encountered in telecommunication terminal devices or accessories. Fifteen acoustic set ups were used, each driven with a moderate voltage (within the specified range for the transducer) and with a high voltage (exceeding the specified range for the transducer). The over-driving resulted in strong nonlinear distortion. This gave thirty “real” systems with both linear and nonlinear distortion.

Figure 10 is a scatter plot of obtained versus predicted ratings for music stimuli with artificial combined distor-

tions (asterisks) and real combined distortions (open circles). The model predicted the ratings well for both real and artificial distortions. The correlation between obtained and predicted ratings was 0.90. The corresponding correlation for speech stimuli was 0.85.

7. DISCUSSION

The model for the perceptual effects of nonlinear distortion is based on cross-correlation of the outputs of simulated auditory filters in response to the original and distorted signals. The output of each filter can be considered as a relatively slowly varying envelope (ENV) imposed on a rapidly oscillating carrier (the temporal fine structure, TFS) [21]. The cross-correlation accommodates the effects of the distortion on both ENV and TFS. An alternative approach is based on the changes in ENV, which can be modelled using an array of simulated auditory filters followed by an array of simulated modulation filters [22]. This forms the basis for a model that combines the power-spectrum model of masking [16] with the model of modulation processing of Dau *et al.* [22]. The combined model is called the “generalized power spectrum model” [23]. A version of this model has been shown to be accurate in predicting the perceptual effects of a large range of distortions [24]. However, Biberger *et al.* [24] stated that the model is probably “not able to account for quality degradations typically occurring in telecommunication as a result of phase jitter, variable delay, or time warping.” The model described in Sects. 5 and 6 of this

paper may be more accurate for these types of distortion. This requires further investigation.

8. CONCLUSIONS

The perceptual effects of linear distortion can be predicted accurately using a model based on differences in the excitation patterns evoked by the original and distorted signals. The perceptual effects of nonlinear distortion can be predicted accurately using a model based on differences in the waveforms at the outputs of an array of simulated auditory filters in response to the original and distorted signals. The combined effects of linear and nonlinear distortion can be predicted using a weighted sum of predicted ratings from the linear and nonlinear models, with greater weight for nonlinear distortion.

ACKNOWLEDGEMENTS

I thank Chin-Tuan Tan for his collaboration in the work reported here. Contributions were also made by Nick Zacharov, Ville-Veikko Mattila, and Brian Glasberg. I also thank Chin-Tuan Tan, Tom Baer and Hedwig Gockel for helpful comments on an earlier version of this paper.

REFERENCES

- [1] B. C. J. Moore and C. T. Tan, "Perceived naturalness of spectrally distorted speech and music," *J. Acoust. Soc. Am.*, **114**, 408–419 (2003).
- [2] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th ed. (Brill, Leiden, The Netherlands, 2012).
- [3] B. C. J. Moore, B. R. Glasberg and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, **45**, 224–240 (1997).
- [4] B. C. J. Moore, M. A. Stone and J. I. Alcántara, "Comparison of the electroacoustic characteristics of five hearing aids," *Br. J. Audiol.*, **35**, 307–325 (2001).
- [5] E. Zwicker, "Die elementaren Grundlagen zur Bestimmung der Informationskapazität des Gehörs (The foundations for determining the information capacity of the auditory system)," *Acustica*, **6**, 356–381 (1956).
- [6] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, **74**, 750–753 (1983).
- [7] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, **12**, 47–65 (1940).
- [8] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, **47**, 103–138 (1990).
- [9] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," *J. Acoust. Soc. Am.*, **59**, 640–654 (1976).
- [10] T. Houtgast, *Lateral Suppression in Hearing* (Ph.D. Thesis, Free University of Amsterdam, 1974).
- [11] B. C. J. Moore and B. J. O'Loughlin, "The use of non-simultaneous masking to measure frequency selectivity and suppression," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. (Academic, London, 1986), pp. 179–250.
- [12] A. J. Oxenham and C. A. Shera, "Estimates of human cochlear tuning at low levels using forward and simultaneous masking," *J. Assoc. Res. Otolaryngol.*, **4**, 541–554 (2003).
- [13] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Hetu, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. El Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, R. Meredith, T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman and C. Ludvigson, "An international comparison of long-term average speech spectra," *J. Acoust. Soc. Am.*, **96**, 2108–2120 (1994).
- [14] C. T. Tan, B. C. J. Moore and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *J. Audio Eng. Soc.*, **51**, 1012–1031 (2003).
- [15] C. T. Tan, B. C. J. Moore, N. Zacharov and V.-V. Matilla, "Predicting the perceived quality of nonlinearly distorted music and speech signals," *J. Audio Eng. Soc.*, **52**, 699–711 (2004).
- [16] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, B. C. J. Moore, Ed. (Academic, London, 1986), pp. 123–177.
- [17] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *J. Audio Eng. Soc.*, **50**, 331–342 (2002).
- [18] R. D. Patterson, M. H. Allerhand and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.*, **98**, 1890–1894 (1995).
- [19] C. Jurado, C. S. Pedersen and B. C. J. Moore, "Psychophysical tuning curves for frequencies below 100 Hz," *J. Acoust. Soc. Am.*, **129**, 3166–3180 (2011).
- [20] B. C. J. Moore, C. T. Tan, N. Zacharov and V.-V. Matilla, "Measuring and predicting the perceived quality of music and speech subjected to combined linear and nonlinear distortion," *J. Audio Eng. Soc.*, **52**, 1228–1244 (2004).
- [21] B. C. J. Moore, *Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss* (World Scientific, Singapore, 2014).
- [22] T. Dau, B. Kollmeier and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," *J. Acoust. Soc. Am.*, **102**, 2906–2919 (1997).
- [23] T. Biberger and S. D. Ewert, "Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility," *J. Acoust. Soc. Am.*, **140**, 1023–1038 (2016).
- [24] T. Biberger, J.-H. Flessner, R. Huber and S. D. Ewert, "An objective audio quality measure based on power and envelope power cues," *J. Audio Eng. Soc.*, **66**, 578–593 (2018).