Perceived naturalness of spectrally distorted speech and music

Brian C. J. Moore and Chin-Tuan Tan

Citation: The Journal of the Acoustical Society of America **114**, 408 (2003); doi: 10.1121/1.1577552
View online: https://doi.org/10.1121/1.1577552
View Table of Contents: https://asa.scitation.org/toc/jas/114/1
Published by the Acoustical Society of America

Perceived naturalness of spectrally distorted speech and music

Brian C. J. Moore[a] and Chin-Tuan Tan
Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, England

(Received 30 January 2002; revised 4 April 2003; accepted 7 April 2003)

We determined how the perceived naturalness of music and speech (male and female talkers) signals was affected by various forms of linear filtering, some of which were intended to mimic the spectral ''distortions'' introduced by transducers such as microphones, loudspeakers, and earphones. The filters introduced spectral tilts and ripples of various types, variations in upper and lower cutoff frequency, and combinations of these. All of the differently filtered signals (168 conditions) were intermixed in random order within one block of trials. Levels were adjusted to give approximately equal loudness in all conditions. Listeners were required to judge the perceptual quality (naturalness) of the filtered signals on a scale from 1 to 10. For spectral ripples, perceived quality decreased with increasing ripple density up to 0.2 ripple/ERB_N and with increasing ripple depth. Spectral tilts also degraded quality, and the effects were similar for positive and negative tilts. Ripples and/or tilts degraded quality more when they extended over a wide frequency range (87–6981 Hz) than when they extended over subranges. Low- and mid-frequency ranges were roughly equally important for music, but the mid-range was most important for speech. For music, the highest quality was obtained for the broadband signal (55–16 854 Hz). Increasing the lower cutoff frequency from 55 Hz resulted in a clear degradation of quality. There was also a distinct degradation as the upper cutoff frequency was decreased from 16 845 Hz. For speech, there was a marked degradation when the lower cutoff frequency was increased from 123 to 208 Hz and when the upper cutoff frequency was decreased from 10 869 Hz. Typical telephone bandwidth (313 to 3547 Hz) gave very poor quality.
[DOI: 10.1121/1.1577552]

PACS numbers: 43.66.Lj, 43.38.Md, 43.58.Ry [NFV]

I. INTRODUCTION

It is widely accepted that the perceived quality of signals reproduced by transducers such as loudspeakers or headphones can be strongly affected by the frequency response of the transducers (Toole, 1986a, b; Toole and Olive, 1988). It is also generally accepted that ''smooth'' wideband frequency responses are desirable (Bucklein, 1962; Toole, 1986a, b; Toole and Olive, 1988; Gabrielsson *et al.*, 1990, 1991). However, there have been few studies that have aimed to quantify the relationship between perceived quality and frequency response irregularity and/or bandwidth. This study was designed to provide such information.

Reviews of studies on the perceptual effects of irregularities in frequency response are provided in Toole (1986a, b) and Toole and Olive (1988) and we give only a brief overview here. There have been two general approaches to this issue. Some researchers have used filters to introduce spectral distortions with well-controlled characteristics. Bucklein (1962) investigated the audibility (detectability) of single peaks and dips in frequency response, as a function of the sharpness of the peaks and dips, quantified in terms of their ''Q'' value. He found, using speech and music signals, that broad (low Q) peaks and dips were more easily detected than narrow peaks and dips. Fryer (1977) and Toole and Olive (1988) also studied the effects of single peaks (resonances). Like Bucklein, they found that low-Q resonances were more easily heard than high-Q resonances. They also found that resonances were detected more easily using pink or white noise as the test signal than using speech or music. Toole and Olive (1988) reported that, for impulsive or transient sounds, the addition of reverberation could increase the audibility of medium- and low-Q resonances. Gabrielsson *et al.* (1990) investigated the effect on perceived sound quality of filtering signals in three ways, so as to boost low, medium, or high frequencies by 15–20 dB. A reference ''flat'' response was also used. Speech, music, and noise signals were reproduced via an earphone, and listeners were asked to rate the quality of the reproduced sound in terms of loudness, clarity, spaciousness, brightness, softness/gentleness, nearness, and fidelity. When the reproduction level was relatively high, all of the conditions with boosts in response received lower fidelity ratings than the ''flat'' response condition. However, other results suggest that a mild emphasis of medium to high frequencies can actually lead to increased fidelity ratings (Gabrielsson *et al.*, 1988). Bech (2002) studied the effect of low-frequency cutoff and slope and of ripples in frequency response in the low-frequency range (up to 120 Hz). Subjects judged the magnitude of upper bass and lower bass relative to a fixed reference system, using various pieces of pop music as the signal. The cutoff frequency had a significant effect on the ratings of lower bass and the amount of ripple affected ratings of both lower and upper bass. Studies of this type have not attempted to quan-

[a]Electronic mail: bcjm@cus.cam.ac.uk

tify the relationship between frequency response (smoothness or bandwidth) and perceived fidelity.

Other researchers have relied on variations in frequency response from one loudspeaker to another to explore how frequency response affects perceived sound quality. Gabrielsson *et al.* (1991) obtained subjective ratings of musical sounds reproduced by 18 different high-fidelity loudspeakers. They advanced several hypotheses about the relationship between specific perceptual scales and frequency response. For example, ''clarity'' was hypothesized to be ''favored by a broad frequency range and a smooth response with a certain emphasis on midhigh frequencies,'' while ''fullness'' was hypothesized to be ''favored by a broad frequency range but with more emphasis on lower frequencies.'' They found that their hypotheses were supported better when the loudspeaker frequency responses were measured in the listening room than when they were measured in free field or in a reverberation room. However, no quantitative relationships were established between the frequency-response shapes and the perceptual ratings.

Toole (1986a, b) used natural variations between loudspeakers to examine the relationship between frequency responses and listener preferences. He reported that ''it is possible to see a progressive increase in the smoothness as a function of fidelity ratings. The loudspeakers with lower ratings tend to exhibit more fine structure in the curves which, in the spatially averaged data, indicates the presence of resonances within the loudspeaker system.'' He also emphasized the importance of the directivity pattern of the loudspeakers, and suggested that a good loudspeaker should have as uniform a directivity as possible over a wide frequency range. Again, however, no attempt was made to quantify the relationship between the degree of smoothness of the frequency response and the preferences of the listeners.

There are several limitations of previous studies that have used natural variations in loudspeakers to explore the perceptual consequences of nonflat frequency responses. First, the studies have mostly used high-fidelity loudspeakers, so the deviations from flat responses have been relatively small, typically less than ± 4 dB. Thus, the results are not applicable to many of the transducers that are in common use, such as those in telephones, low-cost earphones, portable radios, public address systems, and computers. Second, the frequency response of loudspeakers can be measured in many different ways (for example, on-axis in an anechoic room, averaged for many directions in an anechoic room, in a reverberation chamber, in typical listening rooms) and there is no clear consensus as to the ''best'' method. Third, loudspeakers can differ in many factors other than their frequency response (e.g., in nonlinear distortion, phase response, and directivity) and it is difficult to isolate the effects of these other factors from the effects of differences in frequency response. Finally, these studies have not provided quantitative estimates of the effect on perceived sound quality of different types of spectral distortions.

The ability to detect changes in the spectral shape of stimuli has been widely studied in the context of profile analysis (Green, 1988). However, studies of this topic have generally required subjects to distinguish between two stimuli differing in spectral shape (and level), and, with a few exceptions (Gockel and Colonius, 1997), the spectral characteristics of the stimuli have been fixed throughout a ''run.'' This is not representative of the situation of listening to speech or music through a transducer with a nonflat frequency response, where the signal itself has a spectrum that changes over time. Also, studies of profile analysis have mainly involved the measurement of thresholds for detection of a change in spectral shape, rather than quantification of the perceptual effect of a change in spectral shape.

The goal of the present study was to explore how the perceived naturalness of music and speech signals was affected by various forms of linear filtering. We wanted to isolate the effects of irregularities in amplitude response from other forms of distortion that can occur in transducers, such as phase distortion and nonlinear distortion. Of course, the overall quality of a transducer is determined by the combined effects of all such factors. Our ultimate goal is to produce a perceptually based model that can predict the effects of all forms of distortion. However, we believe that this can only be achieved by studying the effect of each form of distortion separately. The data presented here are intended to represent a first step towards characterizing the perceptual effect of nonflat amplitude responses.

Some of the types of filtering that we introduced were intended to mimic the spectral distortions that can occur in transducers. These included spectral tilts and band-limiting. We also included spectral ripples that were sinusoidal on an ERB_N scale (see below for details). Although ripples in the frequency response of conventional loudspeakers are not usually regular, some transducers, such as the distributed mode (flat panel) loudspeaker, do show fairly regular ripples in their response (Gontcharov *et al.*, 1999). The artificial regular ripples used here were intended to allow us to characterize the effects of ripple density and center frequency. 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 a high-quality earphone with a good approximation to a diffuse field characteristic, which was especially selected for its low harmonic and intermodulation distortion. This allowed us to be confident that the main factor affecting the subjective responses was the spectral filtering. However, because no reverberation was present in the filtered stimuli, the results should not be considered as applicable to sounds reproduced via loudspeakers in a normal listening room; rather, the results are applicable to sound reproduction via headphones and earpieces, such as those used in telephones. An important feature of our approach was that all filtering conditions were randomly intermixed. This helped to ensure that the response criteria of the subjects were applied consistently to all the various forms of spectral distortion.

II. METHOD

A. Filtering conditions

For many applications, such as in the forthcoming standard for high-quality telephony, the frequency range of inter-

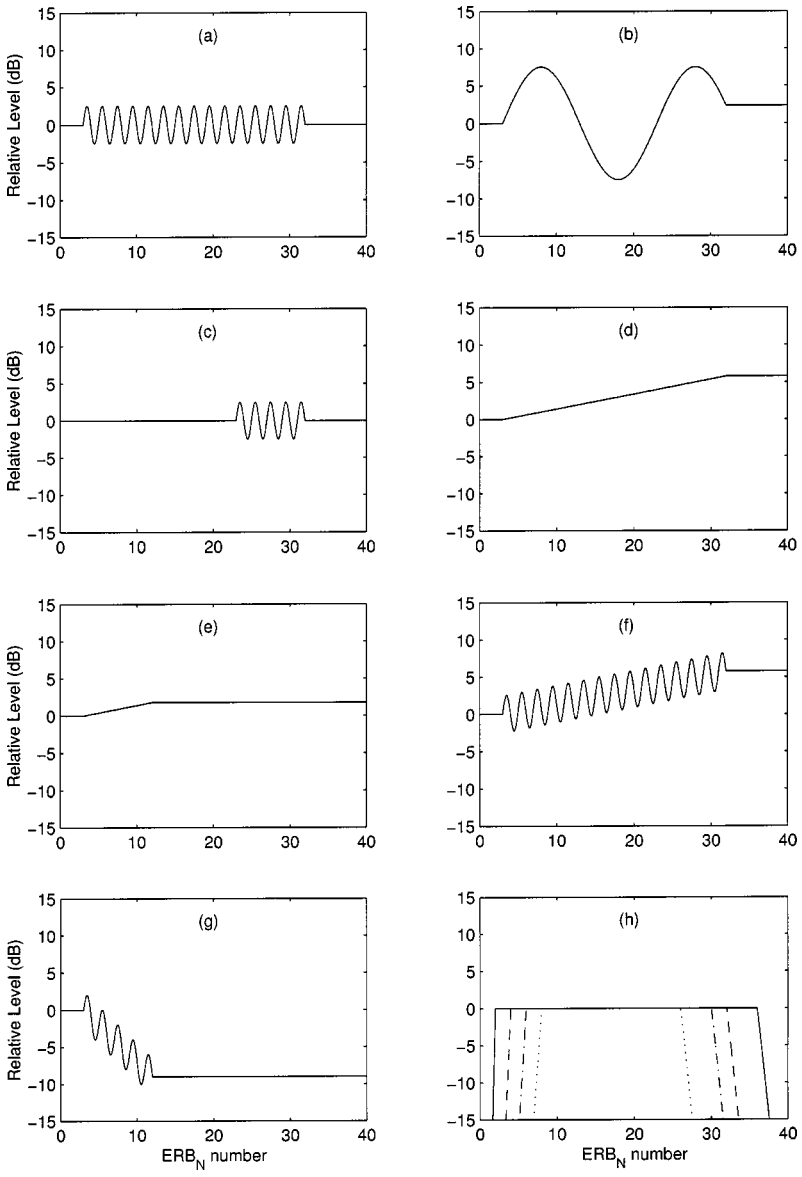


FIG. 1. Schematic illustration of some of the filters used to introduce spectral distortions. The abscissa shows frequency expressed as ERB_N number. (a) Spectral ripple with 5-dB depth and 0.5 ripples/ERB_N, over the range 3–32 ERB_N. (b) Spectral ripple with 1.5-dB depth and 0.05 ripples/ERB_N, over the range 3–32 ERB_N. (c) Spectral ripple with 5-dB depth and 0.5 ripples/ERB_N, over the range 23–32 ERB_N. (d) Positive spectral tilt of 0.2 dB/ERB_N, over the range 3–32 ERB_N. (e) Positive spectral tilt of 0.2 dB/ERB_N, over the range 3–12 ERB_N. (f) Spectral ripple with 5-dB depth and 0.5 ripples/ERB_N, combined with positive spectral tilt of 0.2 dB/ERB_N, both over the range 3–32 ERB_N. (g) Spectral ripple with 5-dB depth and 0.5 ripples/ERB_N, combined with negative spectral tilt of -1 dB/ERB_N, both over the range 3–12 ERB_N. (h) Band-limiting with lower cutoff frequency of 2, 4, 6, or 8 ERB_N and upper cutoff frequency of 26, 30, 32 or 36 ERB_N.

est extends from about 100 to 7000 Hz (frequencies outside that range are not transmitted). As we wished our results to be applicable to telephony, for the set of conditions (1)–(3) below, the spectral distortions were restricted approximately to this frequency range, although the stimuli had a wide bandwidth. We included other conditions with bandpass filtering, including some that simulate current telephone standards (typically 300 to 3400 Hz) and some that present broadband stimuli. In what follows, the characteristics of the filters are described with frequency transformed to ERB_N number, where ERB_N stands for the equivalent rectangular bandwidth of the auditory filter for normal-hearing listeners, when measured at a moderate sound level. The ERB_N scale is a perceptually relevant scale (Moore and Glasberg, 1983; Glasberg and Moore, 1990; Moore, 2003), expressed by the following equation:

$$\text{ERB}_\text{N}\ \text{number} = 21.4 \log_{10}(4.37F + 1), \quad (1)$$

where F is frequency in kHz. This scale is conceptually similar to the Bark scale proposed by Zwicker and co-workers (Zwicker and Terhardt, 1980), although it differs somewhat in numerical values.

Response-modifying filters with the following shapes were used:

(1) Spectral ripples that were sinusoidal in dB on an ERB_N scale. Ripple rates of 0.05, 0.1, 0.2, and 0.5 ripples/ERB_N were used. Ripple depths (peak-to-valley ratios in dB) were 5, 10, and 15 dB. The ripples always started (at the low-frequency end of the range) at a ''zero-crossing.'' The ripples either extended over the range 3–32 ERB_N (87 to 6981 Hz), or were limited to a restricted frequency range, with a flat response outside that range. An example of a wide-range ripple with rate 0.5 ripple/ERB_N and depth 5 dB is shown in Fig. 1(a). In this case, the ripple ended (at the high-frequency end) at a zero crossing. An example of a wide-range ripple with rate 0.05 ripple/ERB_N and depth 15 dB is shown in Fig. 1(b). In this case the ripple ended slightly away from a zero crossing, as only 1.45 ripples were present. The filter responses for frequencies above the ripple range

remained at the value set by the end of the ripple (possible consequences of this are discussed later). Three subranges were used: 3–12, 13–22, and 23–32 ERB_N (corresponding to frequency ranges of 87–606, 701–2224, and 2503–6981 Hz). An example of a ripple with rate 0.5 ripples per ERB_N and depth 5 dB, restricted to the range 23–32 ERB_N, is shown in Fig. 1(c). The use of these subranges allowed us to determine if the perceptual effect of spectral ripples is relatively constant across frequency or varies with frequency. The total number of conditions with ripple was 4 (ripple rates)×3 (ripple depths)×4 (ranges)=48.

(2) Spectral tilts that were linear in dB/ERB_N. Tilts of ±0.1, 0.2, 0.5, and 1 dB/ERB_N were used. The tilts either extended over the frequency range 87 to 6981 Hz, or were limited to a restricted frequency range, with a flat response outside that range. An example of a wide-range positive tilt of 0.2 dB/ERB_N is shown in Fig. 1(d). Three subranges were used: 3–12, 13–22, and 23–32 ERB_N. An example of a positive tilt of 0.2 dB/ERB_N in the subrange 3–12 ERB_N is shown in Fig. 1(e). It should be noted that, for a given tilt in dB/ERB_N, the difference in level between the two ends of the tilt region was three times as great for the wide-range tilt as for any subrange tilt. The total number of conditions with tilt was 8 (tilts)×4 (ranges)=32.

(3) Spectral ripples [as described in (1)] combined with spectral tilts [as described in (2)]. Ripples with rates of 0.1 and 0.5 dB/ERB_N and depths of 5 and 10 dB were combined with tilts of ±0.2 and ±1 dB/ERB_N. Again, the ripples and tilts either extended over the frequency range 87 to 6981 Hz, or were limited to one of the three subranges. An example of full range ripple and a positive tilt of 0.2 dB/ERB_N is given in Fig. 1(f). An example of a ripple combined with a negative tilt of −1 dB/ERB_N, both restricted to the subrange 3–12 ERB_N, is shown in Fig. 1(g). The total number of conditions with both tilt and ripple was 2 (ripple rates)×2 (ripple depths)×4 (tilts)×4 (ranges)=64.

(4) Bandpass filtering with all possible combinations of the following lower and upper cutoff frequencies: lower, 2, 4, 6, and 8 ERB_N (55, 123, 208, and 313 Hz); upper 26, 28, 30, 32, 36, and 40 ERB_N (3547, 4455, 5583, 6981, 10 869, and 16 854 Hz). The response was flat within the passband. Examples of the filter shapes for this condition are given in Fig. 1(h). This gave 24 conditions in total.

B. Filter implementation and stimulus presentation

Each of the 168 filters was implemented digitally, using the overlap-add method (Allen, 1977), which does not introduce any phase distortion. Digital representations of the input signals were obtained directly from CD, using the standard sampling rate of 44 100 Hz. Frames containing 4096 samples (duration =92.9 ms) were Hanning windowed and overlapped by 75%. Thus, frames were updated every 23.2 ms. Two sets of stimuli were used, speech and music, which were evaluated in separate testing sessions. The speech was a concatenation of two sentences, one from a male and one from a female talker, taken from tracks 49 and 50 of the CD ''Sound Quality Assessment Material'' (SQAM) produced by the European Broadcasting Union (www.ebu.ch: materials also available from http://sound.media.mit.edu/mpeg4/audio/sqam/). The same two sentences were used throughout. The overall duration of the two sentences, including the brief pause between them, was 3.1 s. The music was a fragment of jazz (piano, bass and drums) with a relatively constant overall level, taken from a commercial CD (digital recording). The same fragment was used throughout. Its duration was 7.3 s. The speech and music were filtered off-line and the filtered stimuli were stored on computer disk.

The upper panels in Fig. 2 show the power spectral density for the speech (left) and music (right), averaged over the whole duration of the samples. The spectral density is greatest at low frequencies. The lower panels show excitation patterns calculated from the spectra using the method described by Moore *et al.* (1997), and plotted as a function of ERB_N number; the equivalent frequency in Hz is plotted at the top of each panel. This method takes into account the free-field to eardrum transformation (Shaw, 1974), which boosts frequencies around 3 kHz, and the effect of the middle ear, which results in a reduction in the relative level of very low and very high frequencies (Puria *et al.*, 1997; Aibara *et al.*, 2001). The excitation patterns are somewhat flatter than the power spectra, and for both speech and music the peak excitation occurs at an ERB_N number of about 10 (equivalent to about 440 Hz).

The stimuli were replayed to the listener using a 24-bit Lynx 1 sound card, mounted in a PC. The output of the sound card drove Sennheiser HD580 earphones. The same signal was fed to each earpiece. These earphones have a diffuse field response, i.e., they produce at the eardrum of the listener a similar frequency as would be obtained listening in a diffuse sound field. Thus their response at the eardrum shows an increase in the frequency range around 3000 Hz, which reflects the resonance normally produced by the concha and meatus. The earphones were calibrated using a KEMAR manikin (Burkhard and Sachs, 1975), averaging the results for the ''large'' and ''small'' ears. The output of the ear stimulator was connected to a Hewlett-Packard 35670A dynamic signal analyzer. The frequency response was compared to mean of the diffuse field responses of the human ear measured by Shaw (1974), Kuhn (1979), and Killion *et al.* (1987). The response of the earphone was found to match the mean diffuse field response within ±3.5 dB from 30 to 6000 Hz. Above 6000 Hz, the response showed some irregularities which varied depending on which ear was used in KEMAR and also depending on the exact positioning of the earphones on the manikin. Additional measurements using a probe microphone (Etymõtic Research ER7C) close to the eardrum of several human individuals showed that the response above 6000 Hz varied from one individual to another, but that the response averaged across individuals was close to the diffuse field response. Such individual variations also occur in the diffuse field responses of humans (Shaw, 1974; Kuhn, 1979; Killion *et al.*, 1987).

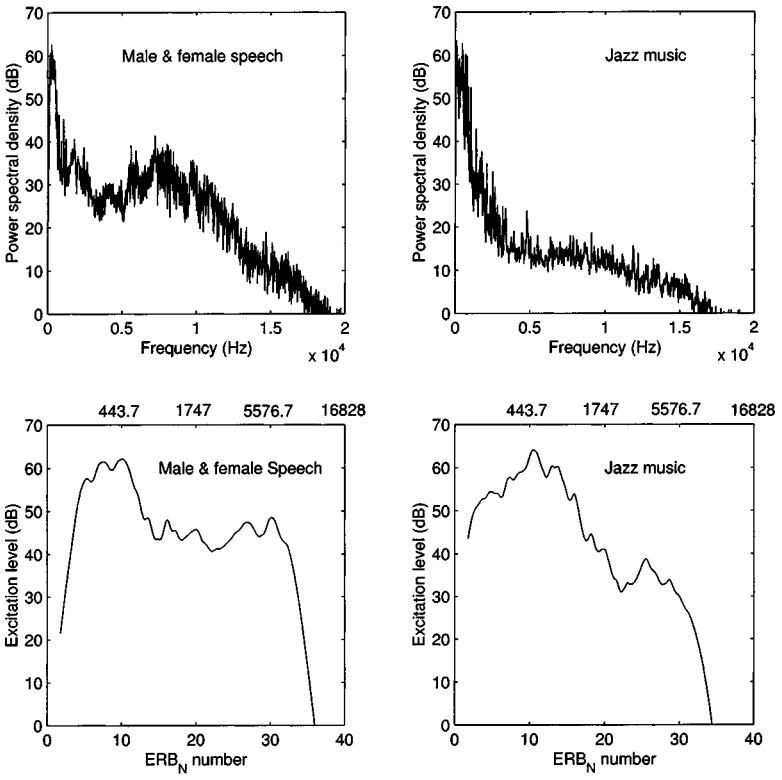


FIG. 2. The upper panels show the mean power spectral density as a function of linear frequency for the speech (left) and music (right). The lower panels show corresponding excitation patterns plotted as a function of ERB_N number.

We also measured the harmonic and intermodulation distortion produced by the earphone. For sinusoidal inputs with frequencies from 100 to 6000 Hz, and for sound levels at the output up to 90 dB SPL, the level of the distortion component corresponding to the second harmonic was always at least 70 dB below the level of the primary component, while the level of the distortion component corresponding to the third harmonic was always at least 84 dB down. Other distortion components were not measurable. For various pairs of primary frequencies, f_1 and f_2, in the range 100 to 7000 Hz, and for levels up to 90 dB SPL, the level of the distortion component at f_2-f_1 was always unmeasurable. For the same conditions, the level of the distortion component at $2f_1-f_2$ was always at least 73 dB lower in level than the primary tones, and was usually unmeasurable. No other distortion components were measurable. We conclude that the nonlinear distortion produced by the earphones was probably below the audible limit (Jacobs and Wittman, 1964; Letowski, 1975; Gabrielsson *et al.*, 1976).

The overall level of each filtered signal was adjusted digitally prior to digital–analog conversion so as to give roughly a constant loudness of 86.4 phons (binaural listening). The required adjustment was calculated using the loudness model of Moore *et al.* (1997). The calculations took into account the diffuse field response of the earphones.

C. Experimental method

In a given session, a listener was tested using either speech or music stimuli. Filtered stimuli were presented in a randomized order. After each stimulus presentation, there was a pause, during which the listener 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 response categories were displayed on the computer screen, and subjects responded using the mouse to ''click'' on their category of choice. The computer waited indefinitely until a response was made. The next stimulus was presented approximately 1 s after a response was made.

To illustrate the meaning of the descriptors for the categories, before the experiment proper started, samples were presented of wideband signals (55 to 16 854 Hz) without any spectral ripple or tilt; these were described as examples of category 10. Similarly, samples were presented with large amounts of ripple and spectral tilt and described as examples of category 1. In the experiment proper, each subject was tested in four sessions on different days, twice using speech and twice using music. The repeated measurement for each type of stimulus allowed us to assess how consistent the responses of each subject were. Half the subjects were tested first using speech and half using music. In a given session the 168 different conditions were intermixed and presented in random order. Subjects were allowed to pause and rest at any time during a session. A session typically lasted about 1 h for the speech stimuli and 1 h and 10 min for the music stimuli.

D. Subjects

Ten subjects were tested. None had a history of hearing disorders and all had audiometric thresholds better than or equal to 20 dB HL in both ears at all audiometric frequencies from 250 to 8000 Hz. The mean audiometric threshold across subjects and ears was better than 7.3 dB for all audiometric frequencies from 250 to 8000 Hz. Their ages ranged from 15 to 31 years (mean 24, standard deviation 6.2). Subjects were paid for their participation.

TABLE I. Correlation of the mean ratings across sessions for each individual subject with the mean ratings across subjects. Correlations are given separately for each subject for the music stimuli and the speech stimuli. A high correlation for a given subject indicates that the pattern of results for that subject is similar to the pattern found in the mean results.

Subject	1	2	3	4	5	6	7	8	9	10
Music	0.94	0.90	0.90	0.88	0.90	0.89	0.78	0.89	0.85	0.82
Speech	0.92	0.93	0.91	0.84	0.93	0.94	0.93	0.94	0.92	0.87

III. RESULTS

The results for each subject generally showed a very similar pattern across the two test sessions for a given type of signal (speech or music). The overall consistency across test sessions was assessed by calculating the mean score across subjects for each condition and stimulus type, separately for each session, and then calculating the correlation of the scores for the 168 conditions across sessions. The correlations obtained in this way were 0.97 for the music stimuli and 0.97 for the speech stimuli. The very high correlations indicate a high degree of consistency across test sessions.

The pattern of results was also very consistent across subjects. To assess the degree of consistency across subjects, we calculated the mean score for each subject and each condition across the two sessions. We also calculated the mean score across subjects for each condition and stimulus type, including the data for both sessions. Then, for each subject in turn, we calculated the correlation between the scores for that individual subject and the mean scores, over the 168 conditions. The higher the correlation, the more closely the pattern of scores for a given subject resembles that for the group as a whole. The resulting correlations are shown in Table I, separately for music and for speech stimuli. The correlations are generally high, all but one being above 0.82. The high correlations indicate a high degree of consistency across subjects. Hence, in what follows, we focus on the mean results. For simplicity of presentation, most of the figures show the effect of only one type of spectral manipulation.

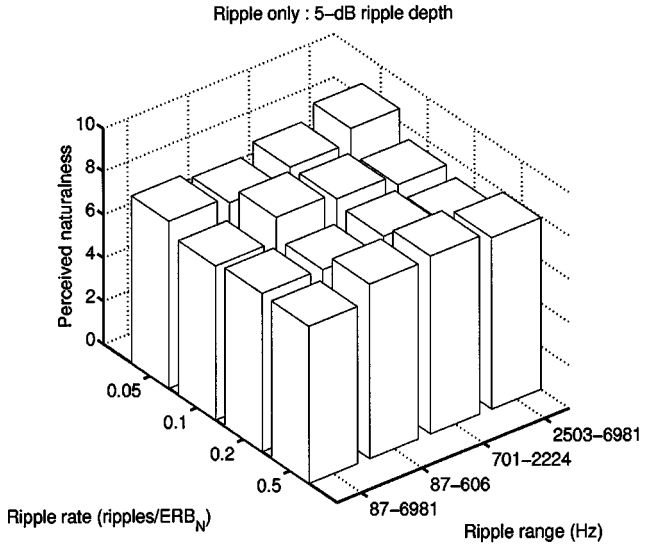
A. Spectral ripples

Figure 3 shows the mean ratings for music (top) and speech (bottom) when processed through filters giving spectral ripples with a depth of 5 dB (peak-valley ratio). Ripples of this magnitude occur quite commonly in the frequency response of good-quality transducers. The *x* axis shows the ripple rate (ripples/ERB_N), the *y* axis shows the ripple range, and the *z* axis shows the mean rating.

Spectral ripples with a depth of 5 dB have only a moderate effect on perceived naturalness. The ripples degrade naturalness more when they extend over a wide frequency range (87–6981 Hz) than when they extend over subranges. Perceived naturalness tends to decrease with increasing ripple rate up to 0.2 ripples/ERB_N.

Figure 4 is similar to Fig. 3, but shows results for a ripple depth of 10 dB. This is the kind of spectral ripple that might be found in the response of lower-quality transducers, such as those in medium quality headphones. The degradation in naturalness produced by this greater ripple depth is clearly evident. Again, the ripples degrade naturalness more when they extend over a wide frequency range (87–6981 Hz) than when they extend over subranges, and perceived naturalness tends to decrease with increasing ripple rate up to 0.2 ripples/ERB_N.

Figure 5 is similar to Fig. 3, but shows results for a ripple depth of 15 dB. This is the kind of spectral ripple that might be found in the response of poor quality earpieces and headphones. This large ripple depth produces substantial degradations of naturalness, and for the wide-range ripple the



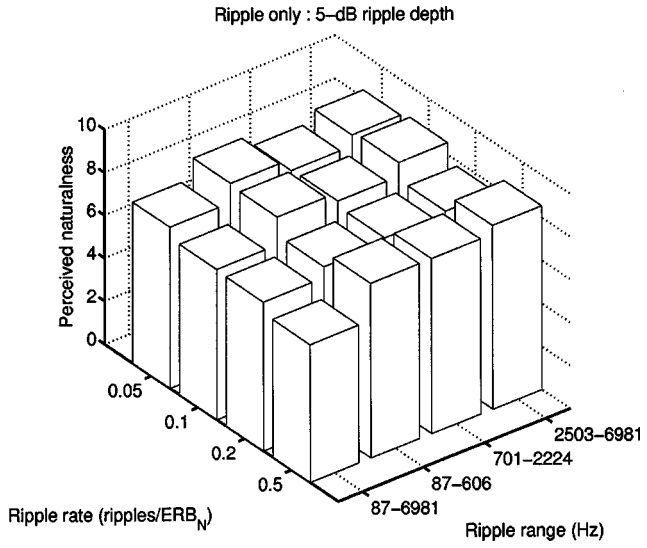
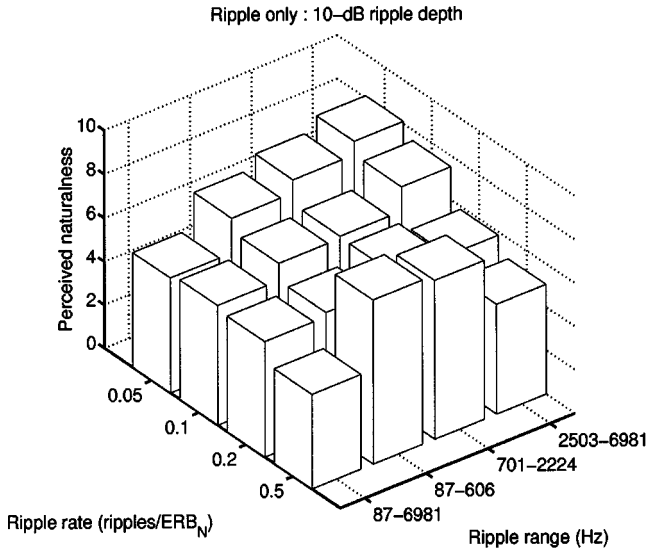
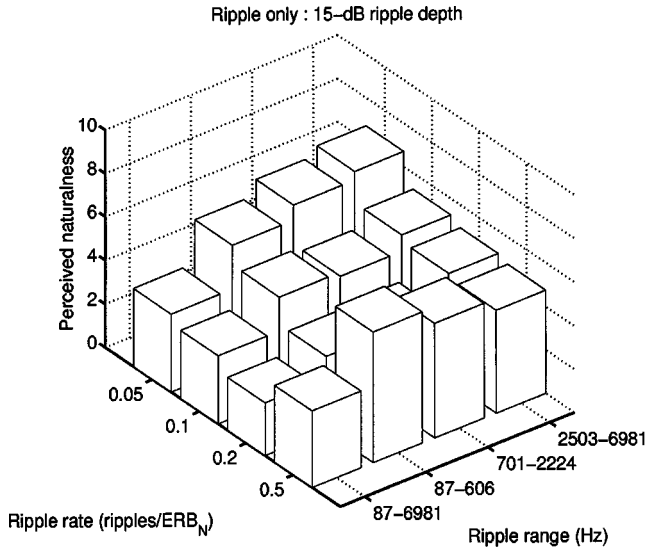


FIG. 3. Mean ratings for music (top) and speech (bottom) when processed through filters giving spectral ripples with a depth of 5 dB. The *x* axis shows the ripple rate (ripples/ERB_N), the *y* axis shows the ripple range, and the *z* axis shows the mean rating.





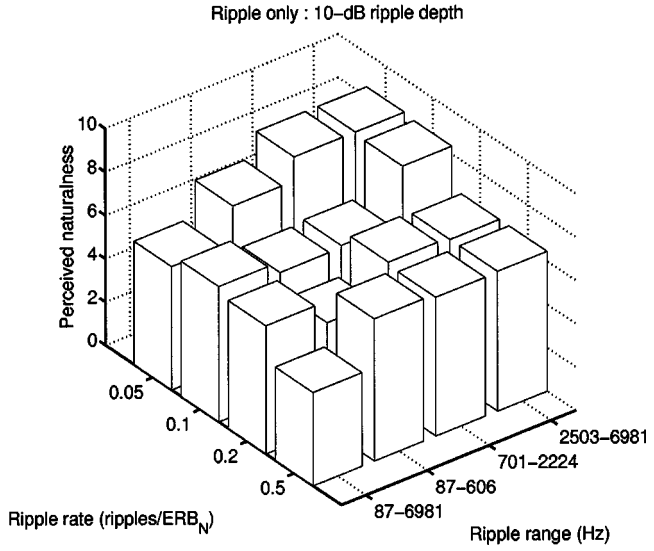


FIG. 4. As Fig. 3, but for a ripple depth of 10 dB.

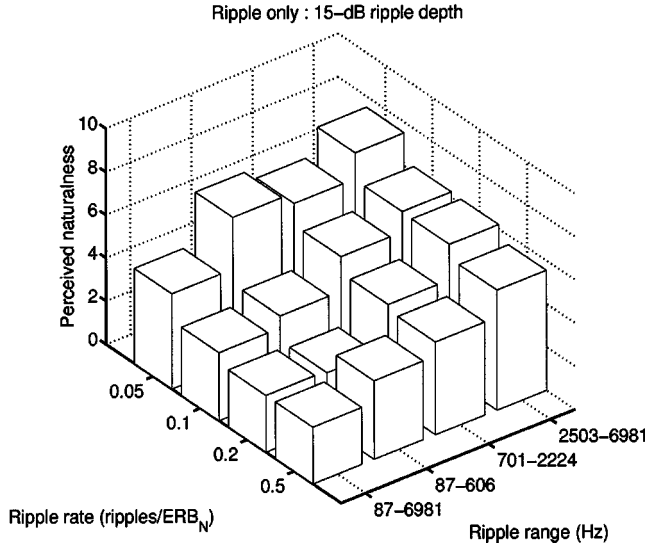


FIG. 5. As Fig. 3, but for a ripple depth of 15 dB.

ratings decrease to values between 2 and 3 for the higher ripple densities. For the speech signal, ripple over the highest frequency subrange produces less degradation than ripple over the low- and mid-frequency subranges.

To assess the statistical significance of the observed effects, the data were subjected to within-subject analyses of variance (ANOVAs). The data from the two test sessions for a given type of stimulus (speech or music) were treated as replications. The least-significant differences test (Snedecor and Cochran, 1967) was used to perform *posthoc* comparisons of means for specific conditions. We performed separate ANOVAs for the music and speech signals, each with factors ripple depth (three values), ripple rate (four values), and ripple range (four values).

For both music and speech, all main effects were significant at $p<0.001$. For music, the mean ratings were 7.7, 6.2, and 4.6 for ripple depths of 5, 10, and 15 dB, respectively. For speech, the corresponding ratings were 7.8, 6.4, and 4.7. For both speech and music, all pairwise comparisons were significant at $p<0.001$. For music, the mean ratings were 6.8, 6.1, 5.5, and 6.2 for ripple rates of 0.05, 0.1, 0.2, and 0.5 ripples/ERB$_N$, respectively. For speech, the corresponding rates were 7.2, 6.3, 5.7, and 5.9. In both cases, the ratings did not differ significantly for rates of 0.1 and 0.5, but all other pairwise comparisons were significant at $p<0.01$. Thus, naturalness at first decreased with increasing ripple rate, but then increased again.

For music, the wide-range ripples gave significantly lower naturalness ratings than all of the subrange ripples ($p<0.001$), but there was no significant effect of subrange center frequency. For speech, wide-range ripples had the greatest deleterious effect, and ripple in subranges produced an effect which decreased with increasing center frequency; the mean rating was 5.2 for the wide-range ripple (3–32 ERB$_N$) and 6.1, 6.6, and 7.3 for the subranges 3–12, 13–22, and 23–32 ERB$_N$, respectively. All pairwise comparisons were significant at $p<0.002$. All of the two-way interactions and the three-way interaction were significant at $p<0.001$.

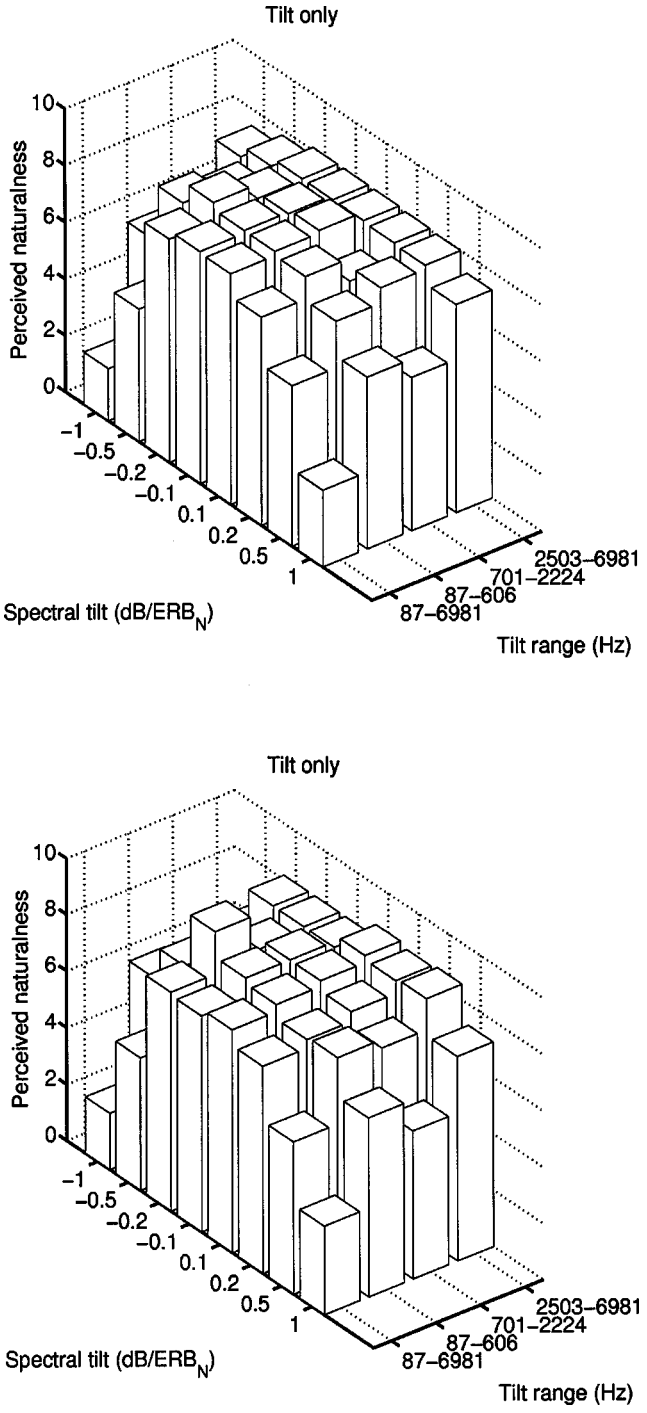


FIG. 6. Mean ratings for music (top) and speech (bottom) with spectral tilts differing in amount and direction and applied over different ranges. The *x* axis shows the spectral tilt (dB/ERB$_N$), the *y* axis shows the tilt range, and the *z* axis shows the mean rating.

However, no interaction accounted for more than 2.5% of the variance in the data, so we will not discuss the interactions further.

B. Spectral tilt

Figure 6 shows results for music (top) and speech (bottom) signals with spectral tilts differing in amount and direction and applied over different ranges. The *x* axis shows the spectral tilt (dB/ERB$_N$), the *y* axis shows the tilt range, and the *z* axis shows the mean rating. Spectral tilts (either upwards or downwards) degrade naturalness, especially when they are applied over the whole frequency range. For the wise-range tilts of large magnitude, the ratings reach values around 2, for both positive and negative tilts. However, tilts restricted to the high-frequency range (2503–6981 Hz) have relatively little perceptual effect, until they are rather extreme.

Separate within-subjects ANOVAs were conducted for the music and speech stimuli, with factors direction of tilt (positive or negative), tilt magnitude, and tilt range. For both music and speech, the main effect of tilt direction was not significant, but the other two main effects were significant at $p<0.001$. For music, the mean rating was 5.8 for the wide-range tilt (3–32 ERB$_N$) and 7.5, 7.2, and 7.8 for the subranges 3–12, 13–22, and 23–32 ERB$_N$, respectively. The mean ratings did not differ significantly for the 3–12 and 13–22 ERB$_N$ subranges. All other pairwise comparisons were significant at $p<0.05$. For speech, the mean rating was 5.7 for the wide-range tilt (3–32 ERB$_N$) and 7.6, 7.1, and 7.9 for the subranges 3–12, 13–22, and 23–32 ERB$_N$, respectively. All pairwise comparisons were significant at $p<0.01$. Thus, mid-range tilt had a slightly perceptual effect than low- or high-range tilt. This greater effect may have occurred partly because, for the mid-range tilt, there were large frequency ranges below and above the tilt region (i.e., below 701 Hz and above 2224 Hz) across which there was a difference in relative level. For the low-range tilt, the ''plateau'' region on the low-frequency side occurred only below 87 Hz, while for the high-range tilt the plateau on the high-frequency side occurred only above 6981 Hz.

C. Combined tilt and ripple

Figure 7 shows the results for speech and music signals with a positive or negative tilt of 1 dB/ERB$_N$ combined with spectral ripples of 10 dB depth. In each panel, the *x* axis shows the ripple rate, the *y* axis shows the ripple range, and the *z* axis shows the mean rating. A large upward or downward tilt combined with a spectral ripple applied over the whole frequency range leads to a severe degradation of naturalness, for both low and high ripple densities; mean ratings approach 2 for these conditions. However, tilt and ripple restricted to the high-frequency range have little perceptual effect. Tilt and ripple in the mid-range tend to have the largest perceptual effect for speech, but for music the low- and mid-ranges are equally important.

Separate within-subjects ANOVAs were conducted for the music and speech stimuli, with factors direction of tilt, tilt magnitude, ripple depth, ripple rate, and range. For both music and speech, the main effect of tilt direction was not significant. Also, the main effect of ripple depth was not significant. Apparently, when a large tilt is present, its effects ''swamp'' those of ripple depth. The effect of ripple rate was not significant, which is consistent with the findings for ripple only for the two rates used here. The main effect of range was significant at $p<0.001$ for both speech and music. For music, the mean ratings were 5.0 for the wide-range tilt and ripple (3–32 ERB$_N$) and 6.9, 6.8, and 7.6 for the subranges 3–12, 13–22, and 23–32 ERB$_N$, respectively. The

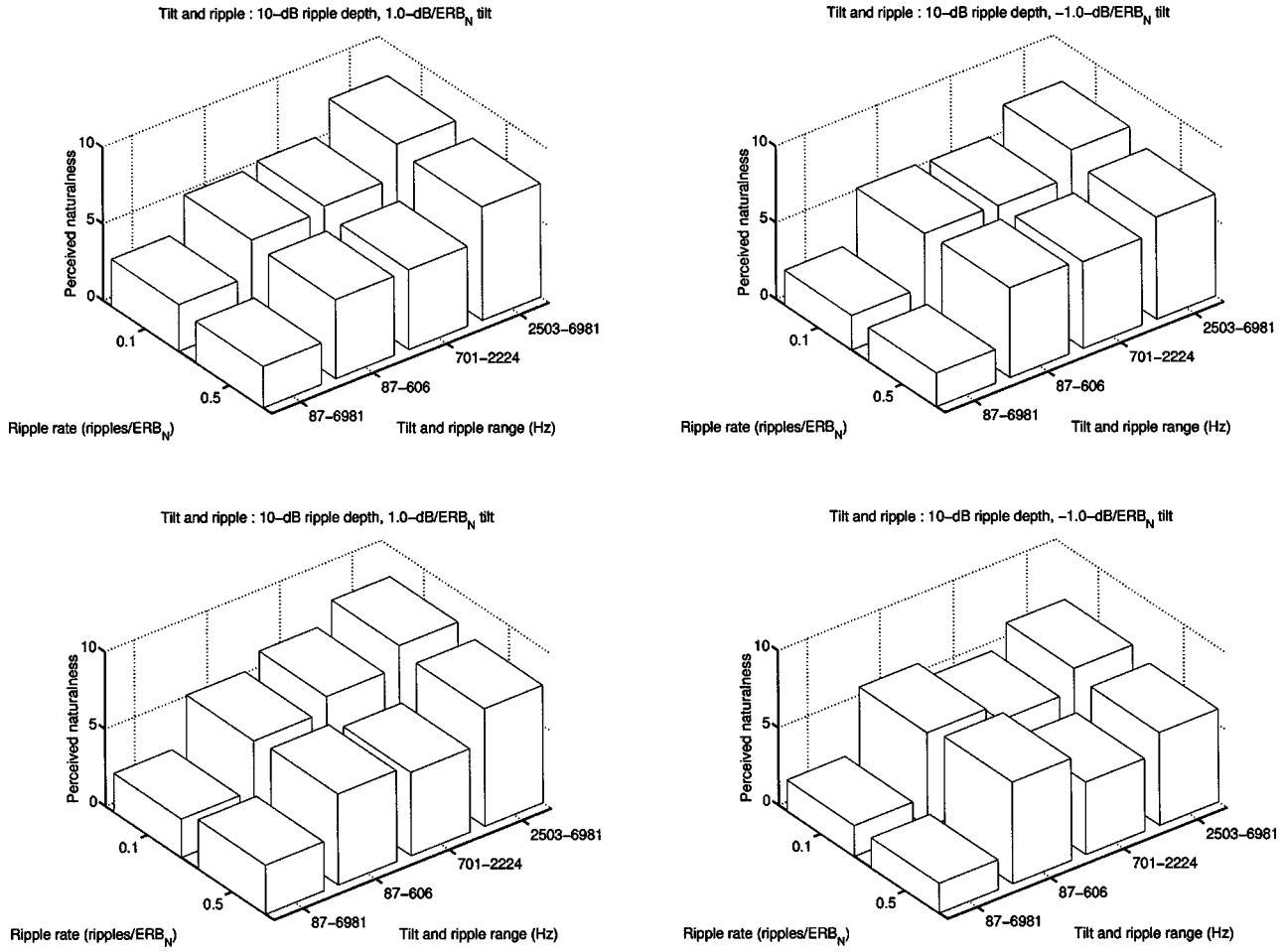


FIG. 7. Mean ratings for music (top) and speech (bottom) with a positive (left) or negative (right) tilt of 1 dB/ERB$_N$ combined with spectral ripples of 10-dB depth. The *x* axis shows the ripple rate, the *y* axis shows the tilt/ripple range, and the *z* axis shows the mean rating.

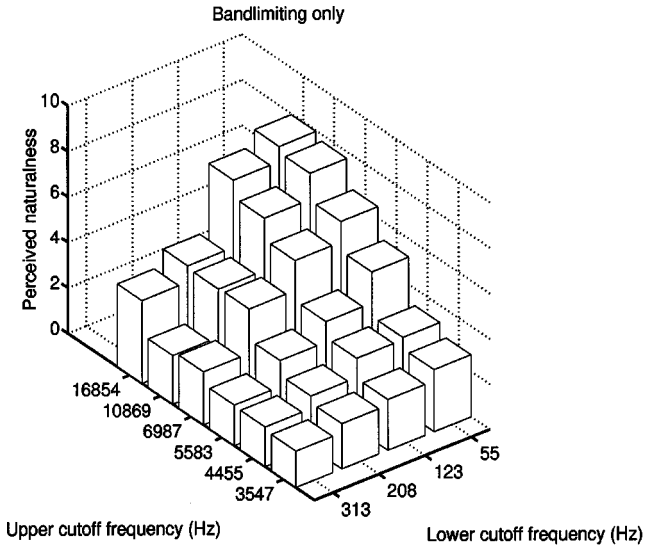
ratings did not differ significantly for ranges of 3–12 and 13–22 ERB$_N$, but all other pairwise comparisons were significant at $p<0.001$. For speech, the mean ratings were 4.8 for the wide-range tilt and ripple and 7.3, 6.8, and 7.7 for the subranges 3–12, 13–22, and 23–32 ERB$_N$, respectively. All pairwise comparisons were significant at $p<0.01$. For both music and speech, the results show that wide-range tilt and ripple has a greater effect than for any subrange, and that tilt and ripple at high frequencies has a smaller effect than at mid or low frequencies.

D. Bandlimiting

Figure 8 illustrates the effects of bandlimiting the stimuli. The *x* axis shows the high-frequency cutoff, the *y* axis shows the low-frequency cutoff, and the *z* axis shows the mean rating. For music (top), the highest naturalness is obtained for the broadband signal (55–16 854 Hz). Increasing the lower cutoff frequency from 55 to 125 Hz results in a clear degradation of naturalness. There is also a reasonably progressive decrease in naturalness as the upper cutoff frequency is decreased from 16 854 to 3547 Hz. For speech (bottom), there is little effect of increasing the lower cutoff frequency from 55 to 123 Hz, but a marked degradation when the lower cutoff frequency is increased from 123 to 208 Hz. Also there is a progressive decrease in naturalness as the upper cutoff frequency is decreased from 10 869 to 3547 Hz. Typical telephone bandwidth (313 to 3547 Hz) gives very poor naturalness for both speech and music.

Within-subjects ANOVAs were conducted separately for the music and speech signals, with factors lower cutoff frequency and upper cutoff frequency. For both music and speech, both main effects and the two-way interaction were significant at $p<0.001$. When the lower cutoff frequency was low (55 or 123 Hz), changes in the upper cutoff frequency had a large effect, while when the lower cutoff frequency was high (313 Hz), changes in the upper cutoff frequency had only a small effect; in the latter case ratings were consistently very low. Similarly, when the upper cutoff frequency was high (10 869 or 18 854 Hz), changes in the lower cutoff frequency had a large effect, while when the upper cutoff frequency was lower (4455 or 3547 Hz), changes in the lower cutoff frequency had only a small effect; again, ratings in the latter case were consistently very low.

The results for both music and speech indicate that poor naturalness introduced by a high lower cutoff frequency cannot be compensated by changing the upper cutoff frequency; and poor naturalness introduced by a low upper cutoff frequency cannot be compensated by changing the lower cutoff frequency.



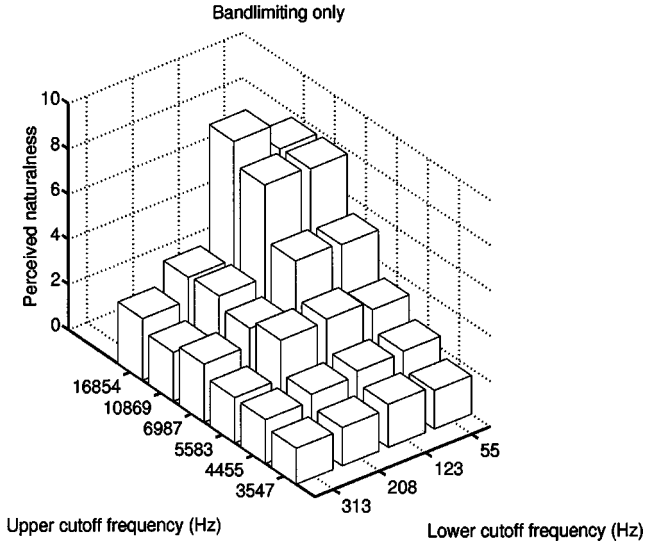


FIG. 8. Mean ratings for band-limited music (top) or speech (bottom) stimuli. The *x* axis shows the high-frequency cutoff, the *y* axis shows the low-frequency cutoff, and the *z* axis shows the mean rating.

IV. DISCUSSION

It is possible that some of the results for the stimuli with spectral ripple were influenced by the fact that the ripple did not always end (on its high-frequency side) at 0 dB; the filter response for frequencies above the ripple range remained at the value set by the end of the ripple. Such an effect was present for all ripple rates except 0.5 ripples/ERB_N. For the 15-dB ripple depth, the "offset" was +2.3, −4.4, and −7.1 dB for the ripple rates of 0.05, 0.1, and 0.2 ripples/ERB_N, respectively. For smaller ripple depths, the offset was correspondingly smaller. The largest offset was somewhat smaller than the −9-dB offset produced by a spectral tilt of −1 dB/ERB_N over a 9-ERB range. While tilts of −1 dB/ERB_N did produce a clear degradation in naturalness (ratings were between 5 and 7, depending on the subrange over which the tilt was applied), the degradation was smaller than produced by 15-dB ripples at a rate of 0.2 ripples/ERB_N, which were in the range 3.5 to 5. Thus, the dominant effect for the ripples was probably due to the ripples *per se*, but the offsets may have played some role.

The effect of the offsets may partly account for the finding that naturalness decreased with increasing ripple density up to 0.2 ripples/ERB_N, but then increased slightly at 0.5 ripples/ERB_N; the ripple rate of 0.5 ripples/ERB_N was the only one for which there was no offset. However, the nonmonotonic effect of ripple rate occurred even when the ripple depth was only 5 dB, when the maximum offset was only −2.4 dB. Thus, some other factor was probably involved. The most likely one is that the limited resolution of the auditory filters has the effect of smoothing fine ripples in the spectrum, so that the representation of ripples in the excitation pattern is reduced.

Our procedure of equating the loudness of all of the filtered stimuli meant that the overall level in different frequency regions differed across stimuli. For example, for stimuli filtered with a positive spectral tilt, the overall level at high frequencies was greater than for the undistorted stimuli, while the level at low frequencies was lower. It is, of course, impossible to introduce differences in spectral shape without introducing differences in overall level in some frequency regions. Subjects may partly have based their judgments on the overall level in specific frequency regions, rather than on the relative spectral shape. However, subjects were specifically asked to judge overall naturalness and not loudness, so we feel that this is unlikely.

Generally, the ratings for the speech stimuli showed a very similar pattern to those for the music stimuli. This is confirmed by the fact that the correlation of the mean ratings for the two stimulus types, taken across all 168 conditions, was 0.96. This high correlation again confirms that the mean ratings are highly reliable and reproducible. We believe that this high reliability is linked to the method that we employed; the use of a random mixture of all conditions, and the inclusion of stimuli with very marked spectral distortion allowed subjects to maintain stable criteria, and helped to give consistent criteria across subjects. It is noteworthy that the mean rating for any condition was never above 8.6 for music and 9.1 for speech, and was never below 1.55 for speech or music. This reflects the well-known reluctance of subjects to use the extremes of the available range of responses when making subjective judgements (Poulton, 1979).

There were some minor differences between the results for the speech and music stimuli. Specifically, for speech, the band-limiting conditions revealed no significant effect of increasing the lower cutoff frequency from 55 to 123 Hz or of decreasing the upper cutoff frequency from 16 854 to 10 869 Hz, while for music the corresponding changes did have significant effects.

There were also some small differences in the relative importance of the different frequency subranges for speech and music. It was always the case, for both speech and music, that wide-range tilt and/or ripple had greater deleterious effects on the naturalness ratings than tilt and/or ripple in any subrange. However, for speech, tilt or tilt plus ripple had a greater effect in the mid-range than in the low range, while

for music the effects were similar for the low- and mid-ranges. For both speech and music, ripple and/or tilt had a smaller effect in the high-range than in the mid- or low-ranges.

It is noteworthy that, for the music stimuli, the mean score with the smallest amounts of tilt (± 0.1 dB/ERB$_N$) was slightly higher at about 8.2 than the mean score in the condition where stimuli were band-limited to the range 55 to 16 854 Hz, but with a flat response in the passband (mean score=7.9). This is probably a consequence of the fact that in the conditions with tilt, the stimuli actually had a wider bandwidth (about 30 to 20 000 Hz) than in the ''broadband'' condition with band-limiting. Probably, frequencies below 55 Hz led to slightly improved naturalness for the music stimuli in the conditions with small amounts of tilt.

Our results provide no support for the claim of Gabrielsson *et al.* (1988) that a mild emphasis of medium to high frequencies can lead to increased fidelity ratings. For both music and speech, scores were almost identical for positive and negative tilts of 0.1 dB/ERB$_N$ over the range 13–22 ERB$_N$ (701–2224 Hz) or 23–32 ERB$_N$ (2503–6981 Hz).

V. CONCLUSIONS

We determined how the perceived naturalness of music and speech signals was affected by various forms of linear filtering, some of which were intended to mimic the spectral ''distortions'' typically introduced by transducers such as microphones, loudspeakers, and earphones. Ratings of naturalness were obtained on a scale from 1 to 10, where 10 represents the most natural and 1 the least. For each stimulus type (music and speech), each of the 168 conditions was rated once in a given test session. Each of ten subjects was tested twice in different sessions for each stimulus type. The main conclusions are as follows:

(1) The ratings were highly consistent across test sessions. The correlation of mean scores across the two sessions was 0.97 for music and 0.97 for speech.
(2) The pattern of scores across conditions (with a few minor exceptions noted below) was very similar for music and speech; the correlation of scores for the two stimulus types was 0.96.
(3) For spectral ripples alone, perceived naturalness decreased progressively as ripple depth was increased. Ratings decreased with increasing ripple rate up to 0.2 ripples/ERB$_N$ and then increased slightly. For both speech and music, the ripples had a greater effect when they occurred over a wide frequency range than when they occurred in subranges. For speech, ripples in the high-frequency range had a smaller effect than ripples in the mid- or low-range.
(4) For tilts alone, ratings decreased progressively with increasing tilt magnitude. The effects were similar for positive and negative tilts. For both speech and music, the tilts had a greater effect when they occurred over a wide frequency range than when they occurred in subranges. This is probably largely a consequence of the fact that, for a fixed tilt value in dB/ERB$_N$, the difference in level between the start and end of the tilt was greater for the wide-range tilt than for the subrange tilts. For both music and speech, tilts in the high-frequency range had a smaller effect than tilts in the mid- or low-range. For speech, tilts in the low-frequency range had a smaller effect than tilts in the mid-frequency range.
(5) For combined tilts and ripples, the degradations in naturalness were greater when the tilts and ripples occurred over a wide frequency range than when they occurred in subranges. For both music and speech, tilts and ripples in the high-frequency range had a smaller effect than tilts in the mid- or low-range.
(6) For band-limited stimuli with a flat response within the passband, there were large effects of manipulating the band edge frequencies.
(7) For music, the highest naturalness was obtained for the broadband signal (55–16 854 Hz). Increasing the lower cutoff frequency from 55 Hz resulted in a clear degradation of naturalness. There was also a distinct degradation as the upper cutoff frequency was decreased from 16 845 Hz. Typical telephone bandwidth (313 to 3547 Hz) gave very poor naturalness.
(8) For speech, there was little effect of increasing the lower cutoff frequency from 55 to 123 Hz, but further increases led to a degradation of naturalness. There was also little effect of decreasing the upper cutoff frequency from 16 854 to 10 869 Hz, but further decreases led to a degradation of naturalness. Typical telephone bandwidth (313 to 3547 Hz) gave very poor naturalness.
(9) For both speech and music, when the lower cutoff frequency was relatively high, changes in the upper cutoff frequency had only a small effect. Similarly, when the upper cutoff frequency was low, changes in the lower cutoff frequency had only a small effect.

ACKNOWLEDGMENTS

This project was supported by Nokia Corporation. We thank Nick Zacharov, Brian Glasberg, and Michael Stone for their help with various aspects of this work. We also thank Michael Stone, Thomas Baer, Neal Viemeister, and three anonymous reviewers for helpful comments on an earlier version of this paper.

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (**2001**). ''Human middle-ear sound transfer function and cochlear input impedance,'' Hear. Res. **152**, 100–109.
Allen, J. B. (**1977**). ''Short term spectral analysis, synthesis and modification by discrete Fourier transform,'' IEEE Trans. Acoust., Speech, Signal Process. **25**, 235–238.
Bech, S. (**2002**). ''Requireents for low-frequency sound reproduction, Part I: The audibility of changes in passband amplitude ripple and lower system cutoff frequency and slope,'' J. Audio Eng. Soc. **50**, 564–580.
Bucklein, R. (**1962**). ''Hörbarkeit von Unregelmässigkeiten in Frequenzgängen bei akustischer Übertragung,'' Frequenz **16**, 103–108.
Burkhard, M. D., and Sachs, R. M. (**1975**). ''Anthropometric manikin for acoustic research,'' J. Acoust. Soc. Am. **58**, 214–222.
Fryer, P. A. (**1977**). ''Loudspeaker distortions: Can we hear them?'' Hi-Fi News Rec. Rev. **22**, 51–56.
Gabrielsson, A., Hagerman, B., Bech-Kristensen, T., and Lundberg, G. (**1990**). ''Perceived sound quality of reproductions with different frequency responses and sound levels,'' J. Acoust. Soc. Am. **88**, 1359–1366.
Gabrielsson, A., Lindström, B., and Till, O. (**1991**). ''Loudspeaker frequency response and perceived sound quality,'' J. Acoust. Soc. Am. **90**, 707–719.

Gabrielsson, A., Nyberg, P. O., Sjögren, H., and Svensson, L. (**1976**). ''Detection of amplitude distortion by normal hearing and hearing impaired subjects,'' Karolinska Institute, Technical Audiology **TA83**, 1–20.
Gabrielsson, A., Schenkman, B. N., and Hagerman, B. (**1988**). ''The effects of different frequency responses on sound quality judgments and speech intelligibility,'' Hear. Res. **31**, 166–177.
Glasberg, B. R., and Moore, B. C. J. (**1990**). ''Derivation of auditory filter shapes from notched-noise data,'' Hear. Res. **47**, 103–138.
Gockel, H., and Colonius, H. (**1997**). ''Auditory profile analysis: Is there perceptual constancy for spectral shape for stimuli roved in frequency?'' J. Acoust. Soc. Am. **102**, 2311–2315.
Gontcharov, V. P., Hill, N., and Taylor, V. (**1999**). ''Measurement Aspects of Distributed Mode Loudspeakers,'' in *AES 106th Convention* (Munich).
Green, D. M. (**1988**). *Profile Analysis* (Oxford U.P., Oxford).
Jacobs, J. E., and Wittman, P. (**1964**). ''Psychoacoustics, the determining factor in stereo disc distortion,'' J. Audio Eng. Soc. **12**, 115–123.
Killion, M. C., Berger, E. H., and Nuss, R. A. (**1987**). ''Diffuse field response of the ear,'' J. Acoust. Soc. Am. Suppl. 1 **81**, S75.
Kuhn, G. (**1979**). ''The pressure transformation from a diffuse field to the external ear and to the body and head surface,'' J. Acoust. Soc. Am. **65**, 991–1000.
Letowski, T. (**1975**). ''Difference limen for nonlinear distortion in sine signals and musical sounds,'' Acustica **34**, 106–110.
Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic, San Diego).
Moore, B. C. J., and 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., Glasberg, B. R., and Baer, T. (**1997**). ''A model for the prediction of thresholds, loudness and partial loudness,'' J. Audio Eng. Soc. **45**, 224–240.
Poulton, E. C. (**1979**). ''Models for the biases in judging sensory magnitude,'' Psychol. Bull. **86**, 777–803.
Puria, S., Rosowski, J. J., and Peake, W. T. (**1997**). ''Sound-pressure measurements in the cochlear vestibule of human-cadaver ears,'' J. Acoust. Soc. Am. **101**, 2754–2770.
Shaw, E. A. G. (**1974**). ''Transformation of sound pressure level from the free field to the eardrum in the horizontal plane,'' J. Acoust. Soc. Am. **56**, 1848–1861.
Snedecor, G. W., and Cochran, W. G. (**1967**). *Statistical Methods* (Iowa U. P., Ames).
Toole, F. E. (**1986a**). ''Loudspeaker measurements and their relationship to listener preferences: Part 1,'' J. Audio Eng. Soc. **34**, 227–235.
Toole, F. E. (**1986b**). ''Loudspeaker measurements and their relationship to listener preferences: Part 2,'' J. Audio Eng. Soc. **34**, 323–348.
Toole, F. E., and Olive, S. E. (**1988**). ''The modification of timbre by resonances: Perception and measurement,'' J. Audio Eng. Soc. **36**, 122–142.
Zwicker, E., and Terhardt, E. (**1980**). ''Analytical expressions for critical band rate and critical bandwidth as a function of frequency,'' J. Acoust. Soc. Am. **68**, 1523–1525.