In perceptual audio coding, it is beneficial to have information about tonality of signals. On one hand, audio codecs handle transient and stationary parts of the signals differently, on the other hand, psychoacoustic models usually distinguish between masking effects of tonal and noise maskers. Asymmetry of masking has been investigated by psychoacousticians to some extent. Recently, we investigated duration thresholds for discrimination of very short tone and narrowband noise bursts. In order to gain more insight, the current study investigates these duration thresholds for the perception of short noise bursts in tones. A pure tone was compared to another tone of same frequency and duration, in which a short segment in the center was replaced by narrowband noise. Two windows (with Hann-like slopes) were used for cross-fading of the middle part. The tests were conducted in 12 conditions, which were the combination of center frequencies 345, 754, 1456, 2658 Hz and bandwidths one ERB (Equivalent Rectangular Bandwidth), ERB/2 and ERB/4. The “pure tone” and the “noise-included tone” (each 400 ms long, sampled at 48 kHz) were randomly ordered with a 400 ms pause in between. Subjects answered the question: “which signal is a pure tone?”. 2-alternative forced choice 3-step up-down method was used for collecting data. The variable was the duration of the noise segment in the center. The up-down step size was chosen to be 5 ms for the first two reversals, 2 ms for the next two reversals and 1 ms for the next (and final) 6 reversals. Only the final 6 reversal points were collected, arithmetic average of which led to the duration threshold for a run. Each condition was tested twice and the results were averaged. 15 normal-hearing subjects participated in the tests. The results for the 12 conditions are shown, and discussed in detail.

1. Introduction

There have been great achievements in the field of perceptual audio coding in the last three decades. One of the interesting areas where there is still a capacity for useful improvements (with respect to perceptual audio codecs) is tonality estimation. Some of the standardized codecs make use of tonality estimation, e.g. in MPEG-4 AAC\textsuperscript{1} [1][2] where transient or stationary segments of the input signals are processed in short or long frames, respectively. Furthermore, in some psychoacoustic models tone and noise maskers are differentiated, since, as known from literature [3][4][5][6], the masking capabilities of tonal and noise maskers are not equal. While tones are weak maskers, noise signals (centered at the same frequencies and of equal energies) mask more strongly. The so called “asymmetry of masking” has been investigated in [5] and [6].

\textsuperscript{1}Moving Picture Experts Group-Standard, Advanced Audio Coding
One aspect of tonality is investigating thresholds of tonality perception. Recently, we carried out some studies for investigating the tonality perception/detection of very short bursts [7][8]. Contrary to those studies, this paper investigates rapid changes in stationary tone bursts, as in natural audio signals, effects such as amplitude and frequency modulation or frequency shifting can also happen for only short segments of them. It means, while in [7] and [8] the bursts’ duration had been changed, in this study, bursts’ duration were constant (400 ms) and only the duration of a middle segment of the signals were changed, which introduced noise-like behavior in the middle of the tone bursts.

In Section 2, the experiment, its stimuli, apparatus, and method are explained in detail. Section 3 deals with the results, statistical analysis, and the limitations of the study. Finally, this paper is summarized in Section 4.

2. The Experiment

Details about the design and conduction of the experiment are listed below.

2.1 Stimuli

In each trial, the stimulus consisted of two bursts and a pause (of 400 ms) in between. One of the bursts was a pure tone of 400 ms played back at 48 kHz sampling rate. The other burst had the same characteristic, however a short segment in the temporal center of the burst was replaced by narrowband noise. Both bursts were switched on and off using a window function with Hann-like slopes (raised cosine) on both sides. Figure 1 shows such a window, an example of the resulting tone burst of 500 Hz and its magnitude spectrum.

![Figure 1. Windowing process for burst generation: (a) window function of 400 ms, (b) a windowed pure tone burst of 500 Hz, and (c) magnitude spectrum of the resulting tone burst (depicted only up to 1000 Hz).](image)

Figure 1. Windowing process for burst generation: (a) window function of 400 ms, (b) a windowed pure tone burst of 500 Hz, and (c) magnitude spectrum of the resulting tone burst (depicted only up to 1000 Hz).

![Figure 2. The middle segment of cross-fading: the solid line shows the window function for eliminating the middle part of the sinusoid; the dashed line shows the window function for introducing narrowband noise to it.](image)

Figure 2. The middle segment of cross-fading: the solid line shows the window function for eliminating the middle part of the sinusoid; the dashed line shows the window function for introducing narrowband noise to it.
Figure 3. A segment of a burst of 400 ms in which the middle part is replaced by narrowband noise of center frequency 500 Hz. On the left and right sides, the burst has the form of a tone of 500 Hz.

The other burst was obtained using the pure tone burst and a long narrowband noise which was generated from bandpass filtered white noise. The center frequency of the narrowband noise corresponded to the frequency of the tone. The middle segment of the pure tone burst was replaced by a segment of the narrowband noise via cross-fading, as depicted in Figure 2. The windows have raised cosine slopes. Up to a noise duration of 5 ms, the length of each cross-fading range was 40% of the length of the entire noise segment. For longer noise segments, the cross-fading were kept constant (2 ms each side). The noise segment that was actually used was determined based on the similarity of the wave forms within the two cross-fading ranges.

Figure 3 illustrates a segment of the noise-included burst. The transition sections are smooth. In the middle, the signal has a narrowband noise shape and on both sides it has a pure tone (500 Hz) characteristic. In the following, we will refer to the duration of the middle noise-segment as the “noise duration”, the transition sections as the “cross-fading ranges”, and the resulting noise-included burst as the “noisy burst”, respectively.

The test series was carried out in 12 conditions, which resulted from the combination of 4 frequencies 345 Hz, 754 Hz, 1456 Hz and 2658 Hz, and 3 bandwidths ERB/4, ERB/2 and ERB$^2$.

2.2 Apparatus

The experiment was carried out with a pair of open headphones, “Sennheiser HD 650”, in an isolated acoustic/studio room with a background noise of approximately 21 dBA$^{SPL}$ (37 dBC$^{SPL}$). By means of a KEMAR artificial head, the system was calibrated for a 250 Hz tone at 65 dBA$^{SPL}$. The frequency response of the headphone was considered and included as an intensity factor for playback. Stimuli were generated in MATLAB on a windows 7 (64 bit) system. The stimuli were regenerated for each trial (e.g. new random narrowband noise for each trial).

2.3 Method

An adaptive 3-step up-down method [9][10] was chosen for the experiment. As shown in Figure 4, subjects were asked to do a 2-AFC task$^3$. As explained before, in each trial, a pair of randomly placed bursts were presented to the subjects (a pure tone burst and a noisy burst). The question was: “Which signal is a pure tone?” As shown in Figure 4 (b) and (c), the subjects received feedback after each trial. Based on correctness of the response of a trial, the course of the experiment was controlled.

\(^2\)ERB: Equivalent Rectangular Bandwidth
\(^3\)2-AFC: Two Alternative Forced Choice
Figure 4. The graphical user interface (GUI) for the experiment: (a) beginning of a run, and feedback after a (b) correct or (c) incorrect response.

by an up-down method. In the following, we will call each test a “run”. Some parameters of the method are described as follows.

- The varying parameter in a run was the noise duration.
- Each run was carried out for 10 reversals, which resulted in 5 minima and 5 maxima. Only data from the last 6 reversals were used for statistical analysis.
- In each run, the step size was 5 ms up to the second reversal point, 2 ms up to the fourth reversal point, and constantly 1 ms up to the final (tenth) reversal point. Since the statistical analysis was done only for the sector of constant step size (1 ms), the threshold could be calculated from the arithmetic mean of the final 6 reversal points. The step size of 1 ms seemed to be adequate [10] according to the standard deviation of the data from previous experiments [8].
- Until the fourth reversal point of a run, a simple up-down method [9][10][11] was used. In doing so, the duration level was brought to the area of the actual threshold of the subject, conveniently and fast. For the remainder of the run, the 3-step up-down method (3-down/1-up) was used [10]. This means that the noise duration was reduced by 1 ms after 3 consecutive correct responses and increased by 1 ms for every incorrect response.
- The starting duration level in each condition had been set before. From the results of [8], as well as from a pilot study with 3 subjects, we had an idea of the placement of the thresholds. The starting points were set distinctly above the hypothetical thresholds, in order to give subjects confidence (and enough adaptation time) for the first couple of the trials. It means, each run started with a few “easy” trials.
- 24 runs were carried out for each subject resulting from 4 frequencies, 3 bandwidths and 2 repetitions: $4 \times 3 \times 2 = 24$.

An example of such a run (of frequency 2658 Hz and bandwidth ERB/4) is depicted in Figure 5. Various parameters of the test can be observed from the figure.

2.4 Subjects

15 subjects participated in the study (average age 28). They were asked whether they had normal hearing, and whether they were currently healthy. Additionally, by means of audiometry, their threshold in quiet was tested as a measure of “normal hearing”. After reading the instructions, they had a training of 10-15 minutes to familiarize with the software, stimuli, and the course of the
Figure 5. An example of a “run”. The parameters of the condition of this run are: frequency of 2658 Hz, bandwidth of ERB/4, and starting noise duration level of 38.5 ms. The vertical dashed lines mark the reversal points. The resulting threshold of 10.33 ms is illustrated with the bold line (magenta). It results from the last 6 reversal points (9.5, 10.5, 9.5, 11.5, 9.5 and 11.5 ms).

experiment. For each subject, the experiment (including the training) took about 85 to 95 minutes, and was carried out on 3 separate days, (30-35 minutes each). After every 3 runs, a short break was scheduled (the GUI was blocked).

3. Results and discussion

As discussed in 2.3, for each of the 24 tests (runs) of a subject, 3 minima and 3 maxima were obtained. The threshold of each subject in that specific run was calculated as the arithmetic mean of these 6 reversal points (e.g. see Figure 5). Since each condition was tested twice, the threshold of a subject in that particular condition was calculated as the average of those two values. The results of the 24 runs of an example subject are listed in Table 1. The results of the “first run” and the “repetition” are the arithmetic mean of the last 6 reversal points of each, respectively. The “threshold” is the average of the two values in each condition.

For each condition, the arithmetic mean of all the thresholds of the individual subjects was calculated as the threshold of that particular condition. Figure 6 illustrates these results. In Figure 6 (a), for each of the 4 frequencies (abscissa), the results of the tests with different noise bandwidths are shown. Ordinate shows the noise duration in ms.

Table 1. All results of an example subject. The noise duration thresholds are given in ms. The values are rounded down.

<table>
<thead>
<tr>
<th>condition</th>
<th>345Hz ERB</th>
<th>754Hz ERB</th>
<th>1456Hz ERB</th>
<th>2658Hz ERB</th>
<th>345Hz ERB/2</th>
<th>754Hz ERB/2</th>
<th>1456Hz ERB/2</th>
<th>2658Hz ERB/2</th>
<th>345Hz ERB/4</th>
<th>754Hz ERB/4</th>
<th>1456Hz ERB/4</th>
<th>2658Hz ERB/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>first run</td>
<td>6.83</td>
<td>4.00</td>
<td>12.66</td>
<td>3.33</td>
<td>8.83</td>
<td>6.83</td>
<td>5.66</td>
<td>7.00</td>
<td>34.00</td>
<td>25.16</td>
<td>15.16</td>
<td>10.16</td>
</tr>
<tr>
<td>repetition</td>
<td>7.83</td>
<td>5.33</td>
<td>5.50</td>
<td>2.83</td>
<td>11.33</td>
<td>8.50</td>
<td>5.33</td>
<td>8.66</td>
<td>29.83</td>
<td>23.16</td>
<td>12.66</td>
<td>8.83</td>
</tr>
<tr>
<td>threshold</td>
<td>7.33</td>
<td>4.66</td>
<td>9.08</td>
<td>3.08</td>
<td>10.08</td>
<td>7.66</td>
<td>5.50</td>
<td>7.83</td>
<td>31.91</td>
<td>24.16</td>
<td>13.91</td>
<td>9.50</td>
</tr>
</tbody>
</table>
Figure 6. Results of psychoacoustic tests of all 15 subjects are illustrated for different conditions. The vertical axes show the noise duration in ms. In (a), the horizontal axis shows frequencies in Hz; for each frequency, the results of tests in different bandwidth conditions are compared to each other. In (b), the horizontal axis shows bandwidths in ERB; for each bandwidth, the results of tests in different frequency conditions are compared to each other. Mean values and standard deviations are depicted in the figures.

Figure 6 (b) illustrates the same results from another perspective. The average and standard deviation of the results of the 4 frequencies are depicted compared to each other in different bandwidth conditions. While abscissa shows the bandwidths, the noise duration is shown along the ordinate.

3.1 Analysis

Analyzing the data shows that increasing frequency or bandwidth reduces the threshold of rapid changes in a tone burst. This agrees with the results of [7] and [8] where the same effect was discovered for tonality detection by investigating very short narrowband noise and tone bursts. However, while the tendencies of the thresholds coincide with the results of [7] and [8], the absolute values of the thresholds of this experiment are somewhat higher. Some explanations are given in the following.

- The questions of the 2-AFC tasks were of different types. While in the two previous studies subjects had been answering whether the stimuli had been identical or different, in this study, they were inquired to find out which stimulus was the pure tone.
- Because of the search for matches for the cross-fading ranges, some part of the noise segments still had entirely tonal characteristics (begin and end of the noise segments, i.e. the cross-fading
ranges). Thus, only a shorter part of the noise segments actually possessed clear noise-like characteristics.

- The need for a cross-fading range also restricted the shape of the noise. E.g., in this study, fewer subjects reported effects such as frequency modulation or frequency shifting (compared to [7] and [8]).

- The stimuli of this experiment were presented at 65 dBA SPL (contrary to 75 dBA SPL in [8]). We believe that the intensities of the signals might also play a role regarding tonality perception/detection. Probably it was more difficult to detect the differences at 65 dBA SPL which led to higher thresholds (longer segments).

3.2 Some notes and limitations

In the following some points and limitations are mentioned.

- As in [8], there seems to be a saturation at low thresholds. E.g., the test could not be easily done accurately for a very high frequency because the resulting threshold would be too low, and the precision of the method would not cover that. As an example, the spectral shape of the signals are always modified by the windows; the shorter the signals (or in this case, the noise duration) the larger the modification.

- An absolute lowest value of 0.5 ms was set in this experiment. Thus, the noise duration could not be less than 0.5 ms. In the course of the study, 1.82% of the trials were in this range. Thus, there was a bias for the values below 1.5 ms (i.e., the lowest step change was from 1.5 to 0.5 ms and reversed), though negligible.

- For cross-fading from tone to noise and vice versa, especially in high frequency conditions, it was not always possible to find an ideal match. Though this might have caused bias, it did not happen often.

4. Conclusions

In this paper, we investigated “the minimum duration of rapid changes in tone bursts” which can be detected by the human auditory system. For generating noisy bursts, the middle segment of the tones were replaced by narrowband noise. Data was collected using a 2-AFC 3-step up-down method. The results show a dependency of the noise duration on bandwidth and center frequency of the noise. This threshold spans mainly from 1 to 45 ms, and decreases with increasing frequency or/and bandwidth (in ERB-scale). Based on these results, we suggest to consider this sensitivity (with respect to frequency and bandwidth) in the implementation and development of psychoacoustic models for perceptual audio codecs.

For audio signal processing applications in cochlear implant devices, as long as they use psychoacoustic models, there might be no need for tonality detection (or equivalently no difference between tone and noise maskers), especially at low frequencies as these devices commonly process the input signal with frames of 8 ms (and most of the thresholds are above this value). In perceptual audio coding, we suggest to apply “frequency and bandwidth sensitive tonality estimation” in psychoacoustic models; e.g., to ensure that, varying temporal resolution should be used for different frequency areas by calculation of the tonality measure.

While this study gave some insight into the dependency of the thresholds on frequency and bandwidth, further investigation of other possible factors (e.g., energy/intensity) is desired.
REFERENCES


